Archives Italiennes de Biologie, 150: 290-326, 2012.

Integrated information theory of consciousness: an updated account

G. TONONI

Department of Psychiatry, University of Wisconsin, Madison, WI, USA

ABSTRACT

This article presents an updated account of integrated information theory of consciousness (IIT) and some of its implications. IIT stems from thought experiments that lead to phenomenological axioms (existence, compositionality, information, integration, exclusion) and corresponding ontological postulates. The information axiom asserts that every experience is specific – it is what it is by differing in its particular way from a large repertoire of alternatives. The integration axiom asserts that each experience is unified – it cannot be reduced to independent components. The exclusion axiom asserts that every experience is definite – it is limited to particular things and not others and flows at a particular speed and resolution. IIT formalizes these intuitions with postulates. The information postulate states that only "differences that make a difference" from the intrinsic perspective of a system matter: a mechanism generates cause-effect information if its present state has selective past causes and selective future effects within a system. The integration postulate states that only information that is irreducible matters: mechanisms generate integrated information only to the extent that the information they generate cannot be partitioned into that generated within independent components. The exclusion postulate states that only maxima of integrated information matter: a mechanism specifies only one maximally irreducible set of past causes and future effects – a concept. A complex is a set of elements specifying a maximally irreducible constellation of concepts, where the maximum is evaluated over elements and at the optimal spatio-temporal scale. Its concepts specify a maximally integrated conceptual information structure or quale, which is identical with an experience. Finally, changes in information integration upon exposure to the environment reflect a system's ability to match the causal structure of the world. After introducing an updated definition of information integration and related quantities, the article presents some theoretical considerations about the relationship between information and causation and about the relational structure of concepts within a quale. It also explores the relationship between the temporal grain size of information integration and the dynamic of metastable states in the corticothalamic complex. Finally, it summarizes how IIT accounts for empirical findings about the neural substrate of consciousness, and how various aspects of phenomenology may in principle be addressed in terms of the geometry of information integration.

Key words

Brain • Experience • Awareness • Causation • Emergence

Phenomenology: Consciousness as integrated information

Everybody knows what consciousness is: it is what vanishes every night when we fall into dreamless sleep and reappears when we wake up or when we dream. Thus, consciousness is synonymous with experience – any experience – of shapes or sounds, thoughts or emotions, about the world or about the self.

It is also common knowledge that our consciousness depends on certain parts of the brain. For example, the widespread destruction of the cerebral cortex leaves people permanently unconscious (vegetative), whereas the complete removal of the cerebellum, even richer in neurons, hardly affects con-

Corresponding Author: Giulio Tononi, 6001 Research Park Boulevard, Madison, WI, 53719, USA - E-mail: gtononi@wisc.edu

sciousness. Furthermore, it matters how the cerebral cortex is functioning. For example, cortical neurons remain active throughout sleep, although their firing patterns may change. Correspondingly, at certain times during sleep consciousness fades, while at other times we dream. It is also well established that different parts of the cortex influence qualitative aspects of consciousness: damage to certain parts of the cortex impairs the experience of color, whereas other lesions impair that of visual shapes. Neuroscientific findings are making progress in identifying the *neural correlates of consciousness* (Koch, 2004). However, to explain *why* experience is generated in the cortex and not in the cerebellum, why it fades in certain stages of sleep, why some cortical areas contribute color and others sound, and to address difficult issues such as the presence and quality of consciousness in newborn babies, in animals, or in pathological conditions, empirical studies are usefully complemented by a theoretical approach. Integrated information theory (IIT) constitutes such an approach. What follows is an outline of IIT, streamlined and updated with respect to previous expositions (Tononi, 2004, 2008).

Three thought experiments

Three thought experiments lie at the heart of IIT – the photodiode thought experiment, the camera thought experiment, and the internet thought experiment.

The photodiode thought experiment

Consider a human and a simple photodiode facing a blank screen that is alternately on and off. The photodiode can tell 'light' from 'dark' just as well as a human. However, a human also has an *experience* of light or dark, whereas the photodiode presumably does not. What is the critical property that humans have and photodiodes lack?

According to IIT, the critical property has to do with how much information is generated when the distinction between light and dark is made. From the intrinsic perspective of a system – photodiode or human – information can best be defined as a "difference that makes a difference"[1]: the more alternatives (differences) can be distinguished, to the extent they lead to distinguishable consequences (make a difference), the greater the information. When the blank screen turns on, the photodiode's mechanism, which can distinguish between a low and a high current, detects a high current and, say, triggers the output 'light' rather than the output 'dark'. Since the distinction is between two alternatives, the photodiode generates 1 bit of information. We take that bit of information to specify 'light' as opposed to 'dark', but it is important to realize that, from the photodiode's perspective, the only specification it can make is whether its input were in one of two ways and whether therefore its outputs should be in one of two ways – this way or not this way. Any further specification is impossible because it does not have mechanisms for it. Therefore, when the photodiode detects and reports 'light', such light cannot possibly mean what it means for us – it does not even mean that it is a visual attribute.

When a human reports pure light, by contrast, mechanisms in his brain distinguish, in a specific way, among a much larger number of alternatives, and are primed accordingly for a large number of different outcomes, thus generating many bits of information. This is because 'light' is distinguished not only from 'dark', but from a multitude of other possibilities, for example a red screen, a green screen, this movie frame, that movie frame, a sound, a different sound, a thought, another thought, and so on. In other words, *each alternative can be distinguished from the others in its own specific way*, and can lead to different consequences, including different verbal reports, actions, thoughts, memories etc. To us, then, 'light' is much more meaningful precisely because we have mechanisms that can specifically distinguish this particular state of affairs we call 'light' against each and every one of a large number of alternatives, and lead to appropriately different consequences. Indeed, as a human, no matter how hard I try, I cannot empty an experience of meaning: I cannot reduce the experience of 'light' to 'this and not this'. More generally, if I am not blind from birth, I cannot reduce myself to lacking visual experiences; if I am not color-blind, I cannot reduce myself to seeing the world in black-and-white; if I know English, I cannot see the word "English" and not understand it; if I am an experienced musician, I cannot reduce myself to listening to a sonata as if I were a novice, and so on.

This central point may be appreciated either by addition or by subtraction. By addition, I realize that I

can only see 'light' the way I see it, as progressively more and more meaning is added by mechanisms that specify how 'light' differs from each of countless alternatives: from various colors, shapes, and countless other visual and non-visual experiences. By subtraction, I can realize that, if I were to lose one neural mechanism after the other, my being conscious of 'light' would degrade – it would lose its non-coloredness, its non-shapedness, it would even lose its visualness – while its meaning is progressively stripped down to just 'one of two ways', as with the photodiode. Either way, the theory says that, the more my mechanisms specify how 'light' differs from its many alternatives, and thereby lead to different consequences – the more they specify what light means – the more I am conscious of it.

The camera thought experiment

Information – the ability to discriminate among a large number of alternatives – is thus an essential ingredient for consciousness. However, another thought experiment, this time involving a digital camera, shows the need for a second ingredient. Assume the sensor chip of the camera is a collection of a million binary photodiodes. Taken together, then, the camera's photodiodes can distinguish among 21,000,000 alternative states, an immense number, corresponding to 1 million bits of information. Indeed, the camera would respond differently to every possible image. Yet few would argue that the camera is conscious. What is the critical difference between a human being and a camera?

According to IIT, the difference has to do with information integration. From the point of view of an external observer, the camera may be considered as a single system with a repertoire of $2^{1,000,000}$ states. However, the chip is not an integrated entity: since its 1 million photodiodes have no way to interact, each photodiode performs its own local discrimination between a low and a high current, completely independent of what every other photodiode might be doing. In reality, the chip is just a collection of 1 million independent photodiodes, each with a repertoire of 2 inputs and outputs – there is no intrinsic point of view associated with the camera chip as a whole. This is easy to see: if the sensor chip were cut into 1 million pieces each holding its individual photodiode, the performance of the camera would not change at all.

By contrast, a human distinguishes among a vast repertoire of alternatives as *a single, integrated system*, one that *cannot be broken down into independent components each with their own separate repertoire*. Phenomenologically, every experience is an integrated whole, one that means what it means by virtue of being one, and which is experienced from a single point of view. For example, no matter how hard I try, experiencing the full visual field cannot be reduced into experiencing separately the left half and the right half. No matter how hard I try, I cannot reduce the experience of a red apple into the separate experience of its color and its shape. Indeed, the only way to split an experience into independent experiences seems to be splitting the brain in two, as in patients who underwent the section of the corpus callosum to treat severe epilepsy (Gazzaniga, 2005). Such patients do indeed experience the left half of the visual field independently of the right side, but then the surgery has created two separate consciousnesses instead of one. Therefore, underlying the unity of experience must be causal interactions among certain elements within the brain. This means that these elements work together as an integrated system, which is why, unlike the camera, their performance breaks down if they are disconnected.

The internet thought experiment

Unlike the camera chip, the internet is obviously integrated – in fact, its main purpose is to permit exchanges of messages between any point of the net and any other point. It can also be used to disseminate or 'broadcast' messages from any one node to many others. The integration is achieved by routers that act as dynamic switches connecting any address in the network with any other address. And yet it seems unlikely that, at least in its current form, the internet is giving rise to some kind of globally integrated consciousness. What could be the critical difference between the network of neurons inside the brain that gives rise to human consciousness, and the network of internet routers connecting devices throughout the world?

According to IIT, the difference has to do with the fact that the neural substrate of consciousness is wired to achieve *maxima of integrated information*, whereas the internet is not. Consider the internet first. The internet is not designed to achieve a maximum of integrated information, but to ensure point to point

communication. Indeed, interactions within the internet can typically be reduced to independent components, and they better be independent, otherwise there would be a chaotic cross-talk and point-to-point communication would not be possible. In other words, the ability to obtain independent, point-to-point signaling excludes the ability to perform global computations, and vice versa. Thus, the internet, while integrated enough to permit point-to-point signaling, is certainly not maximally integrated – not from the intrinsic perspective of the internet itself. On the other hand, from the perspective of an external user, this has great advantages. For example, from a particular node, say the terminal of an information technologist, one can access without any cross-talk a connected hand-held device to diagnose exactly what the speech recognition module is doing or why it may be malfunctioning; or how the power regulating circuits are performing; or one can access a connected peripheral, say a printer, to diagnose if it is running properly; or access anybody else's computer and check any aspect of its functioning; and so on for any other connected device. Moreover, one can check the computations of any connected node at a range of spatial and temporal scales, from the operations performed by individual transistors at microsecond resolution to daily averages of traffic over a hub. However, the price of such complete access is that the internet is not well suited, at least in its current form, to achieve what one may call 'global', autonomous computations.

By contrast, within consciousness information is maximally integrated: every experience is whole, and the entire set of concepts that make up any particular experience – what makes the experience what it is and what it is not – are maximally interrelated. This integration is excellent for a context-dependent understanding of a particular state of affairs, but the flip side of maximal information integration is *exclusion*. No matter how hard I try, I cannot become conscious of what is going on within the modules in my brain that perform language parsing: I hear and understand an English sentence, but I have no conscious access to how the relevant part of my brain are achieving this computation, although of course they must be connected to those other parts that give rise to my present consciousness. Similarly, I have no conscious access to those other parts of my brain that are in charge of blood pressure regulation; or to the complex computations in the cerebellum that help maintain my posture. And I certainly do not have access to whatever is going on in peripheral organs in my body, such as the liver, the kidneys and so on. Furthermore, while I can interact with other people, I have no access to their internal workings. Exclusion applies also within consciousness: at any given time, there is only one consciousness – one maximally integrated subject – me – having one full experience, not a multitude of partial consciousnesses, each experiencing a subset of the contents of my experience. Instead, each experience is compositional, i.e. structured – it is constituted of different aspects in various combinations: I see the shape of the apple, I see its red color, I see a position in space, and I also see that the apple is red and occupies that position. Exclusion also occurs in spatio-temporal terms: what I experience, I experience at a particular spatial and temporal resolution: I have no way to experience directly processes within my brain – even within the parts that are involved in generating experience – that happen at a much finer spatial grain, such as the workings of molecules and atoms within neural cells, or at a much finer temporal grain, such as the millisecond-by-millisecond traffic of spikes among neurons. Similarly, I cannot experience events at a coarser spatial or temporal scale: for example, no matter how hard I try, I cannot lump together into a single experience an entire movie, a waking day, or a lifetime: there is a "right" time scale at which consciousness flows – at other time scales, consciousness simply does not exist.

Phenomenological axioms, ontological postulates, and identities

Based on the intuitions provided by these thought experiments, the main tenets of IIT can be presented as a set of phenomenological axioms, ontological postulates, and identities. The central axioms, which are taken to be immediately evident, are as follows:

An initial axiom is simply that *consciousness exists*. Paraphrasing Descartes, "I experience therefore I am"[2].

Another axiom concerns *compositionality*: *experience is structured*, *consisting of multiple aspects in various combinations*. Thus, even an experience of

pure darkness and silence contains visual and auditory aspects, spatial aspects such as left center and right, and so on.

A central axiom concerns *information*: *experience is informative* or *specific* – in that *it differs in its particular way from other possible experiences*. Thus, an experience of pure darkness and silence is what it is by differing, in its particular way, from an immense number of other possible experiences – including the experiences triggered by any frame of any possible movie.

Another axiom concerns *integration*: *experience is integrated* – in that *it cannot be reduced to independent components*. Thus, experiencing the word "SONO" written in the middle of a blank page cannot be reduced to an experience of the word "SO" at the right border of a half-page, plus an experience of he word "NO" on the left border of another half-page – the experience is whole.

Yet another axiom is *exclusion*: *experience is exclusive* – in that *it has definite borders, temporal, and spatial grain*. Thus, an experience encompasses what it does, and nothing more; at any given time there is only one of its having its full content, it flows at a particular speed, and it has a certain resolution such that certain distinctions are possible and finer or coarser distinctions are not.

To parallel the phenomenological axioms, IIT posits some *ontological postulates*:

An initial postulate is simply that *mechanisms in a state exist*. That is, there are operators that, given an input, produce an output, and at a given time such operators are in a particular state.

Another postulate concerns *compositionality: mechanisms can be structured, forming higher order mechanisms in various combinations*.

A central postulate concerns *information*: *from the intrinsic perspective of a system, a mechanism in a state generates information only if it has* both *selective causes* and *selective effects* within *the system* – that is, the mechanism must constitute "*a difference that makes a difference within the system*". This intrinsic, causal notion of information can be assessed by examining the cause-effect repertoire (CER) specified by a mechanism in a state – the set of past system states that could have been the causes of its present state and the set of future system states that could have been its effects. If a mechanism in a state does not specify either selective causes or selective effects (for example by lacking inputs or outputs), then the mechanism does not generate any cause-effect information (CEI) within the system. Ontologically, the information postulate claims that, from the intrinsic perspective of a system, only differences that make a difference within the system exist.

Another postulate concerns *integration*: *a mechanism in a state generates integrated information only if it cannot be partitioned into independent submechanisms.* That is, the information generated within a system *should be irreducible* to the information generated within independent sub-systems or independent interactions. Integrated information (φ) can be captured by measuring to what extent the information generated by the whole differs from the information generated by its components (minimum information partition MIP). Ontologically, the integration postulate claims that only irreducible interactions exist intrinsically, i.e. in and of themselves.

Yet another postulate concerns *exclusion*: *a mechanism in a state generates integrated information about only one set of causes and effects – the one that is maximally irreducible*. That is, the mechanism can specify only one pair of causes and effects. By a principle of causal parsimony, this is the pair of causes and effects whose partition would produce the greatest loss of information. This maximally irreducible set of causes and effects is called a *concept*. Exclusion can be captured by measuring the maximum of integrated information $_{max}\varphi^{MIP}$ over all possible cause-effect repertoires of the mechanism over the system. Ontologically, the exclusion postulate claims that only maximally irreducible entities exist intrinsically[3].

As will be discussed below, the postulates can be applied to subsets of elements within a system (mechanisms) as well as to systems (sets of concepts). A system of elements that generates cause-effect information (it has concepts), is irreducible (it cannot be

split into mutually independent subsystems), and is a local maximum of irreducibility (in terms of the concepts it generates) over a set of elements and over an optimal spatio-temporal grain of interactions, constitutes a *complex* – a maximally irreducible entity. In this view, only complexes are entities that exist intrinsically, i.e. in and of themselves.

Finally, IIT posits *identities* between phenomenological aspects and informational/causal aspects of systems. The central identity is the following: *an experience is a maximally integrated conceptual information structure*. Said otherwise, an experience is *a "shape" or maximally irreducible constellation of concepts in qualia space* (a *quale*), where qualia space is a space spanned by all possible past and future states of a complex. In this space, concepts are points in the space whose coordinates are the probabilities of past and future states corresponding to maximally irreducible cause-effect repertoires specified by various subsets of elements.

In what follows, the postulates of IIT are briefly illustrated by considering a set of mechanisms (a candidate system of elements). Within the system, the postulates are the first applied to mechanisms in a state, alone or in combination (all subsets), to identify concepts; then the postulates are applied to different systems of elements and the collection of concepts they generate, in order to identify complexes[4].

Information

The information postulate says that *information is a difference that makes a difference from the intrinsic perspective of a system.* This intrinsic, causal[5] notion of information is assessed by considering if the present state of a mechanism can specify both past causes and future effects within the system.

Within a system X, consider a subset of elements S in its present state s[6]. The information s generates about some subset of elements of X in the past (P) is the *effective information* (EI) between P and s:

$$EI\ (P \mid s) = D\ [(P \mid s),\ P^{Hmax}]$$

where *D* indicates the *difference* between two distributions, in this case between the distribution of P states that could have caused s given its present mechanism and state (the *cause repertoire* CR), and the maximum uncertainty (entropy) distribution P^{Hmax}, in which all P outputs are equally likely a priori[7]. Thus, EI(P|s) represents the differences in the past states of P that that can be detected by mechanism S in its present state s. Similarly, *D* between the distribution of F states that would be the effect of 'fixing' mechanism S in its present state s (the *effect repertoire* ER) and the distribution of states of F in which all F inputs are equally likely (F^{Hmax}), is the effective information s generates about future states of F:

$$EI\ (F \mid s) = D\ [(F \mid s),\ F^{Hmax}]$$

Thus, EI(F|s) represents the differences to the future states of F made by mechanism S being in its present state s. Clearly, EI(P|s) > 0 only if past states of P make a difference to s, and EI(F|s) 0 only if s makes a difference to F.

Based on the information postulate, a mechanism in a state (s) generates information from the intrinsic perspective of a system only if it *both* detects differences in the past states of the system *and* it makes a difference to its future states. That is, s generates information only if it has *both* selective causes (EI(P|s) > 0) *and* selective effects (EI(F|s) > 0). The minimum of the two, which represents the 'bottleneck' in the channel between past causes over P and future effects over F as mediated by the mechanism S in its present state s, is called *cause-effect information (CEI):*

$$CEI(P, F \mid s) = \min\ [\ EI\ (P \mid s),\ EI\ (F \mid s)\ ]$$

Clearly, CEI > 0 only if the system's states make a difference to the mechanism, *and* the state of the mechanism makes a difference to the system. Thus an element that monitors the state of the system (say a parity detector), but has no effects on the system, may be relevant from the extrinsic perspective of an observer, but is irrelevant from the intrinsic perspective of the system, as it makes no difference to it. If CEI > 0, the cause and effect repertoires together can be said to specify a *cause-effect repertoire* (CER).

As an example, consider a mechanism A within an isolated system ABC (Fig. 1). The wiring diagram is unfolded into a directed acyclic graph over past, present, and future. A's mechanism is a logical AND

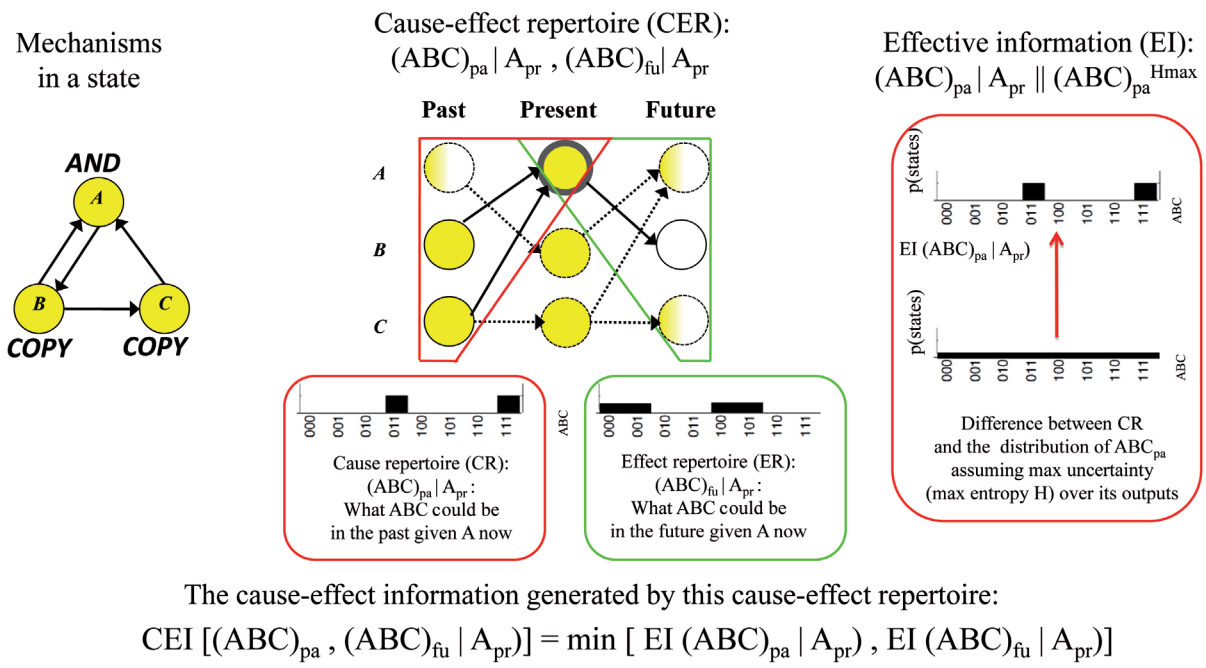


Fig. 1. - A cause-effect repertoire (CER) and the cause-effect information it generates ("differences that make a difference"). See text for explanation.

gate of elements B and C, turning ON if both B and C are ON; moreover, if A is ON, it turns OFF B. Thus, A specifies that, starting from the eight possible past states of elements ABC (maximum entropy distribution), only two past outputs of ABC can lead to A's present state (ON) – those in which B and C are both ON (cause repertoire CR), thereby 'detecting differences' and generating EI. Moreover, A specifies that, starting from maximum entropy over the inputs to ABC, A's present state (ON) can only lead to four future states of ABC – those in which B is OFF (effect repertoire ER), thereby 'making a difference'. Together, CR and ER specify the cause-effect repertoire CER = $(ABC)_{pa} \mid A_{pr}$, $(ABC)_{fu} \mid A_{pr}$ where the subscripts refer to present, past, and future. The cause-effect information (CEI) generated by a mechanism over its cause-effect repertoire (CER) is the minimum between EI $[(ABC)_{pa} \mid A_{pr}]$ and EI $[(ABC)_{fu} \mid A_{pr}]$.

Integration

The integration postulate says that *information is integrated if it cannot be partitioned into independent components*. That is, *a mechanism in state generates integrated information only if it cannot be partitioned into submechanisms with independent causes and effects*. This integrated (irreducible) information is quantified by φ (small phi), a measure of the difference *D* between the repertoire specified by a whole and the product of the repertoires specified by its partition into causally independent components. The difference is taken over the partition that yields the least difference from the whole (the *minimum information partition* (MIP)), i.e. φ^{MIP}[8].

Consider a partition / that splits the interactions between P and S into independent interactions between parts of P and parts of S[9], which can be done by 'injecting' noise (H^{max}) in the connections among them. One can then measure the difference *D* between the unpartitioned cause repertoire CR and the partitioned CR. For the partition that minimizes *D*, known as *minimum information partition* (MIP), the difference *D* is called φ (small phi). The same holds for the difference *D* between the unpartitioned and partitioned effect repertoire ER:

$$\varphi^{MIP} (P \mid s) = D [(P \mid s), \prod (P \mid s / MIP)];$$
$$\varphi^{MIP} (F \mid s) = D [(F \mid s), \prod (F \mid s / MIP)]$$

Thus, φ^{MIP}(P|s) is the 'past' *integrated (irreducible) information*, and φ^{MIP}(F|s) is the 'future' *integrated*

(irreducible) information. Clearly, $\varphi^{MIP}(P|s) > 0$ only if the past states of P make a difference to s that cannot be reduced to differences made by parts of P on parts of s, and likewise for $\varphi^{MIP}(F|s) > 0$.

Based again on the information postulate, a mechanism in a state (s) generates integrated information from the intrinsic perspective of a system only if this information is irreducible *both* in the past *and* in the future. That is, s generates integrated information only if it has *both* irreducible causes ($\varphi^{MIP}(P|s) > 0$) *and* irreducible effects ($\varphi^{MIP}(F|s)>0$). The minimum of the two, which represents the 'bottleneck' in the channel between the past P and the future F as mediated by the mechanism S in its present state s, is called 'cause-effect' *integrated information:*

$$\varphi^{MIP}(P, F \mid s) = \min [\varphi^{MIP}(P \mid s), \varphi^{MIP}(F \mid s)]$$

As an example, Fig. 2a shows a set of 4 elements ABCD, where A is reciprocally connected to B and C is reciprocally connected to D. The wiring diagram is again unfolded into a directed acyclic graph over past, present, and future. Consider now the cause repertoire $(ABCD)_{pa} \mid (ABCD)_{pr}$ and a partition between subsets of elements AB on one side and CD on the other side: $\varphi^{MIP}(P \mid s) = (ABCD)_{pa} \mid (ABCD)_{pr} \parallel (AB)_{pa} \mid (AB)_{pr} \times (CD)_{pa} \mid (CD)_{pr} = 0$. Similarly for the effect repertoire, $\varphi^{MIP}(F \mid s) = (ABCD)_{fu} \mid (ABCD)_{pr} \parallel (AB)_{fu} \mid (AB)_{pr} \times (CD)_{fu} \mid (CD)_{pr} = 0$. Thus, as expected, for this partition $\varphi^{MIP} = \min [\varphi^{MIP}(P \mid s), \varphi^{MIP}(P \mid s)] = 0$. That is, considering the 'whole' CER specified by $(ABCD)_{pa} \mid (ABCD)_{pr}$ and $(ABCD)_{fu} \mid (ABCD)_{pr}$ adds nothing compared to considering the independent 'partial' CER specified by $(AB)_{pa} \mid (AB)_{pr,} (AB)_{fu} \mid (AB)_{pr}$ and by $(CD)_{pa} \mid (CD)_{pr}$, $(CD)_{fu} \mid (CD)_{pr}$. In other words, there is no reason to maintain that the 'whole' CER ABCD exists in and of itself, as it makes no difference above and beyond the two partial CER AB and CD. Thus, searching for partitions among sets of elements yielding $\varphi^{MIP} = 0$ enforces a principle of causal parsimony.

As another example, consider a partition between interactions. The system depicted in Fig. 2b is such that A copies B and B copies A. For the cause-repertoire CR of AB and its partition into independent interactions of A with B and B with A one has that $\varphi^{MIP}(P \mid s) = (AB)_{pa} \mid (AB)_{pr} \parallel (B)_{pa} \mid (A)_{pr} \times (A)_{pa} \mid (B)_{pr} = 0$, and similarly for the effect repertoire ER. That is, the CER of AB over AB (written AB/AB) reduces without loss to the independent CER of A/B

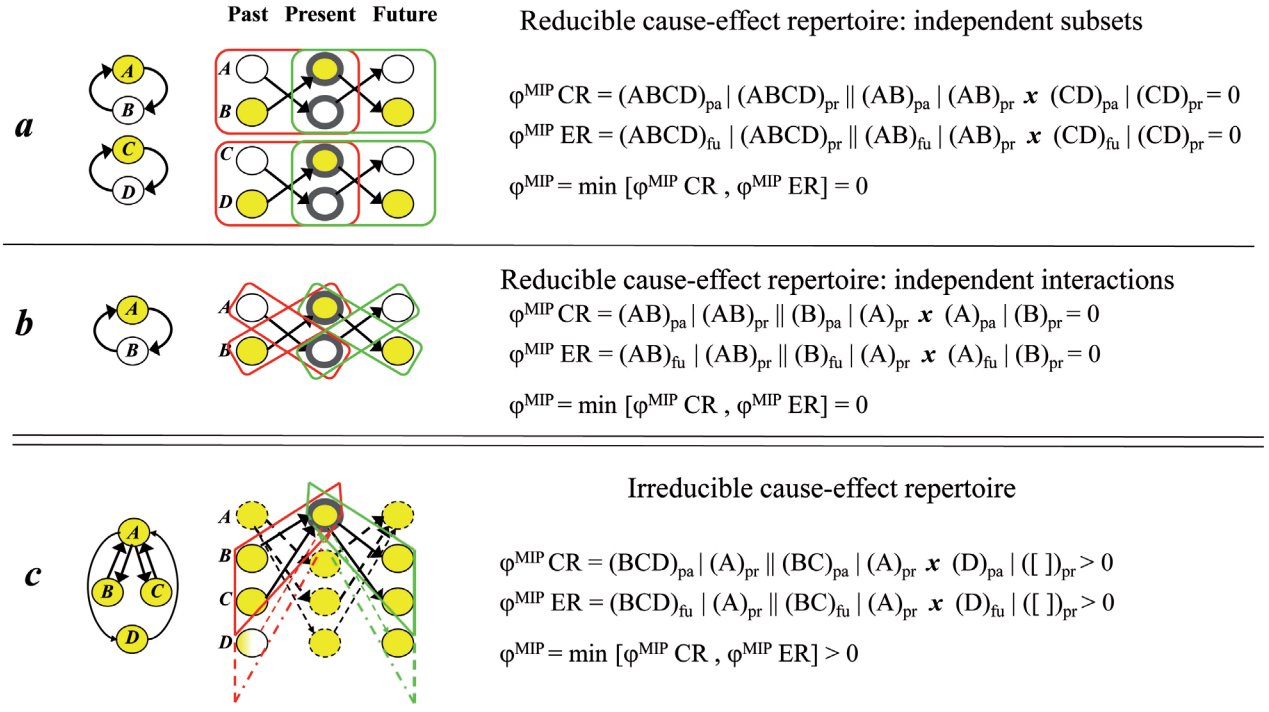


Fig. 2. - Integrated information generated by an irreducible CER, as established by performing partitions. See text for explanation.

and B/A both in the past and in the future. Thus, there is no reason to maintain that the CER AB/AB exists in and of itself, as it makes no difference above and beyond the independent CER of A/B and B/A. Again, searching for partitions among interactions yielding $\varphi^{MIP} = 0$ enforces a principle of causal parsimony.

By contrast, consider a system in which A is a linear threshold unit that receives strong inputs from B and C, which if both ON are sufficient to turn A ON, and a weak input from D; and in which A has strong outputs to B and C (it turns both ON), and a weak output to D (Fig. 2c). Considering the CR of A/BCD, one has that its partition A/BC x D/[] ([] indicates the empty set) yields $\varphi^{MIP} > 0$, and the same holds for the ER. Thus, this CER is irreducible, since there is no way to partition it without losing some information – in this case some information about element D.

Exclusion

The exclusion postulate says that *integrated information is about one set of causes and effects only – those that are maximally irreducible* – other causes and effects are excluded. That is, a mechanism in a state can specify only one pair of causes and effects, which, by a principle of causal parsimony, is the one whose partition would produce the greatest loss of information. This *maximally irreducible set of causes and effects* (MICE) is called a *concept* or, for emphasis, a "*core concept*".

For a given subset of elements S in a present state s, there are potentially many cause repertoires CR depending on the particular subset P one considers (within system X). Exclusion states that, at a given time, s can have only one CR – which is the one having the maximum value of φ^{MIP} ($_{max}\varphi^{MIP}$), where the maximum is taken over all possible subsets P within the system[10]. The corresponding CR is called the *core cause* of s within X. Similarly, the effect repertoire ER having $_{max}\varphi^{MIP}$ over all possible subsets F within the system is called the *core effect* of s within X.

Based again on the information postulate, a mechanism in a state (s) generates integrated information from the intrinsic perspective of a system only if this information is maximally irreducible *both* in the past *and* in the future. That is, s generates maximally integrated information only if it has *both* maximally irreducible causes ($_{max}\varphi^{MIP}(P|s) > 0$) *and* maximally irreducible effects ($_{max}\varphi^{MIP}(F|s) > 0$). The minimum of the two, which represents the 'bottleneck' in the channel between the past P and the future F as mediated by the mechanism S in its present state s, is called 'cause-effect' *maximally integrated information*:

$$_{max}\varphi^{MIP}(P, F \mid s) = \min [_{max}\varphi^{MIP}(P \mid s), _{max}\varphi^{MIP}(F \mid s)]$$

The cause-effect repertoire of s that has $_{max}\varphi^{MIP}(P,F|s)$ within a system X is called a *concept*. Thus, from the intrinsic perspective of a system, a concept is a *maximally irreducible set of causes and effects* (MICE) specified by a mechanism in a state.

For example, in Fig. 3 the powerset of CER (or '*purviews*') of subset A within system ABCD includes, for the cause repertoires, A/A; A/B; A/C; A/D; A/AB; A/AC; A/AD; A/BC; A/BD; A/CD; A/ABC; A/ABD; A/ACD; A/BCD; A/ABCD. Of these, the partition A/BC || A/B x []/C = $_{max}\varphi^{MIP}$ turns out to be maximal (Fig. 3b), higher for example than the partition in Fig. 3a (A/BCD || A/BC x []/D). This is because partitioning away element B (or A) loses much more integrated information than any other partition. A similar result is obtained for the powerset of partitions of A/ABCD for the effect repertoires. By the exclusion postulate, only one CER exists – the one made of the maximally irreducible CR and ER – excluding any other CER[11].

The reason to consider exclusively the CER with $_{max}\varphi^{MIP}$ is as before a principle of causal parsimony – more precisely, a *principle of least reducible reason*. Consider A being ON in the previous example: it specifies a cause repertoire, but cannot distinguish which particular cause was actually responsible for its being ON; and with respect to its effects, it makes no difference which cause turned A ON. Since the particular cause does not matter, the exclusion postulate enforces causal parsimony, defaulting to the maximally irreducible set of causes for A being ON. These least 'dispensable' and thus most likely 'responsible' causes can be called the 'core' causes for A being ON, in the sense that their elimination would have made the most difference[12 13]. In turn, the fact that A is ON also specifies a forward repertoire of possible effects, but once again A should be held most responsible only for its

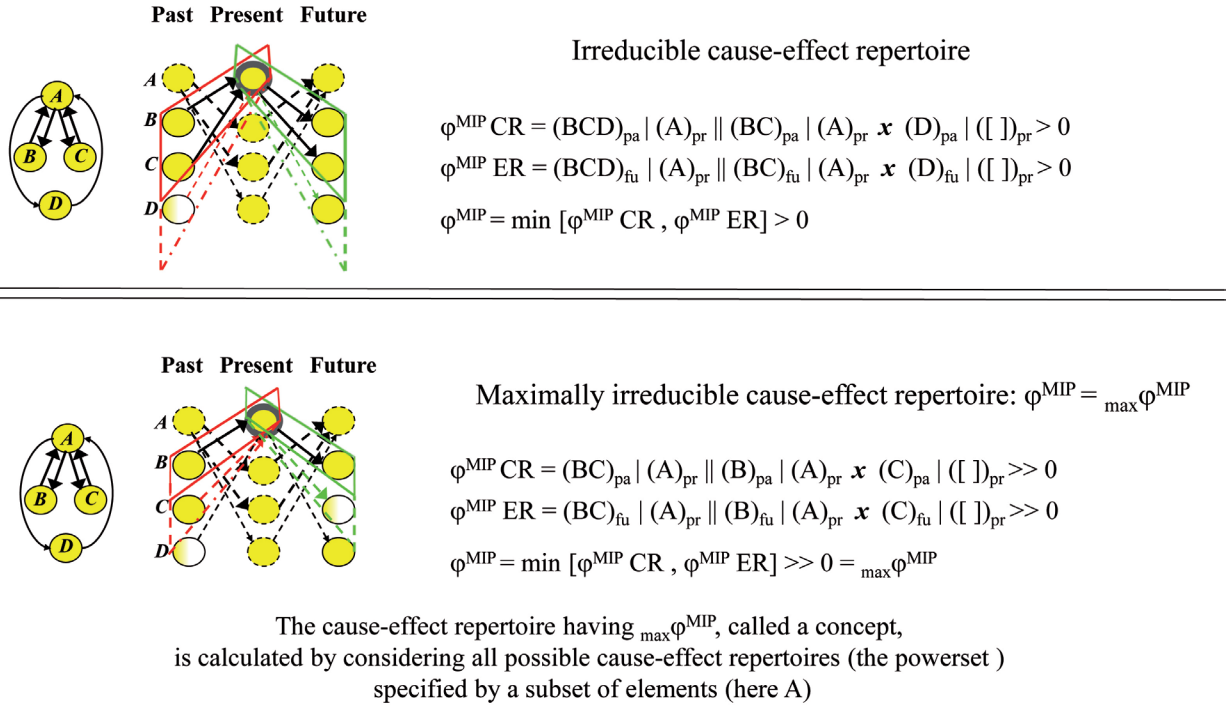


Fig. 3. - Maximally integrated information generated by a maximally irreducible CER over all possible CER specified by a subset of elements within a system. See text for explanation.

maximally irreducible or 'core' effects: the effects for which A being ON is least dispensable, meaning that eliminating A's output would have made the most difference[14].

Concepts

A concept or 'core' concept thus specifies a maximally irreducible cause-effect repertoire (CER) implemented by a mechanism in a state. Within a concept, one can distinguish a core *cause* – the set of past input states (cause repertoire CR) constituting maximally irreducible causes of the present state of the mechanism; and a core *effect* – the set of future output states (effect repertoire ER) constituting maximally irreducible effects of its present state. For example, an element (or set of elements) implementing the concept "table", when ON, specifies 'backward' the maximally irreducible set of inputs that could have caused its turning ON (e.g. seeing, touching, imagining a table); 'forward', it specifies the set of outputs that would be the effects of its turning ON (e.g. thinking of sitting at, writing over, pounding on a table)[15].

As an example, consider the system in Fig. 4, whose wiring diagram is on the left. The middle panel shows the four concepts generated by the system, with their maximally irreducible cause-effect repertoires and the corresponding $_{max}\varphi^{MIP}$. For the concept generated by all three elements (ABC, top row) the figure also shows the product repertoires generated by the minimum information partitions of its maximal cause and effect repertoires.

For a given set of elements, it is useful to consider concepts as points within a space (*concept space*) that has as many axes as the number of possible past and future states of the set (Fig. 4, right panel; the axes are depicted along a circle but should be imagined in a high-dimensional space; the points are indicated as stars). Each concept specifies a maximally irreducible CER, which is a set of probabilities over all possible past and future states, and these probabilities specify a particular point in concept space (more precisely, since probabilities must sum to 1, in the subspace given by the corresponding *concept simplex*). The concept 'exists' with an 'intensity' given by $_{max}\varphi^{MIP}$, that is, its degree of irreducibility (shown by the size of the star).

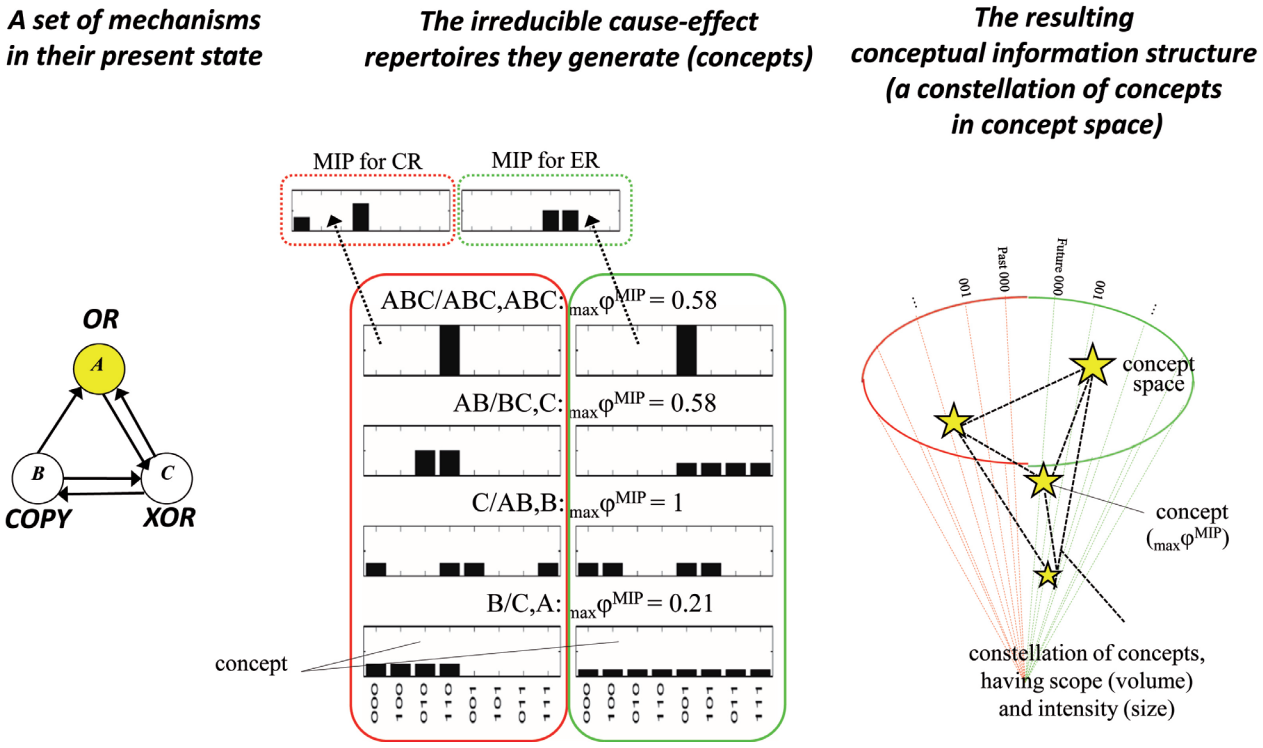


Fig. 4. - An integrated conceptual information structure. See text for explanation.

It is thus possible to evaluate the overall *constellation* of concepts generated by the set of elements in a single concept space, which can be called a *conceptual information structure C*. Among the relevant features one can consider are: i) the intensity, i.e. irreducibility $_{max}\varphi^{MIP}$ of existing concepts; ii) the "shape" of the constellation of concepts in concept space; iii) the dimensionality of the sub-space spanned by all the concepts; iv) the scope of the subspace covered by the concepts; v) the scope of the subspace covered by the concepts weighted by their intensity[16 17].

Complexes

By considering the conceptual information structure C ("constellation" C) specified in concept space by all the concepts generated by a system (Fig. 4), the postulates of IIT can be applied not only to find the maximally irreducible CER of a subset of elements (concepts), but also to find sets of elements, called *complexes*, which generate maximally integrated conceptual information structures. As with concepts, so with complexes, this can be done by: i) making sure, by partitioning the elements of a system, that the constellation of concepts generated by a set of elements cannot be reduced to the product of the constellations generated by the parts (integration postulate); ii) ensuring that the constellation of concepts generated by one part of the system have both selective causes and selective effects in the other part (information postulate); iii) choosing the set of elements that generates the most irreducible constellation of concepts (exclusion postulate).

As before, the irreducibility mandated by the integration postulate can be determined by measuring the difference *D* between the constellation of concepts generated by the whole, unpartitioned set of elements s, and that generated after its partition P into parts:

$$\Phi^{P}\rightarrow (C \mid s) = D\ (C \mid s, C \mid s/P\rightarrow);$$
$$\Phi^{P}\leftarrow (C \mid s) = D\ (C \mid s, C \mid s/P\leftarrow)$$

where the arrow next to P indicates a unidirectional partition, i.e. one that separates causes from effects across the parts by injecting noise in the connections going from one part to the other. Applying as before the information postulate, one has:

$$\Phi^P(C \mid s) = \min [\ \Phi^P{\rightarrow}(C \mid s), \Phi^P{\leftarrow}(C \mid s)\ ]$$

That is, one first partitions across the inputs (causes) to one side of the partition (i.e. the outputs or effects from the other side), then the other way around, and one takes the minimum across the partition. Finally, as before, one finds the partition for which $\Phi^P(C \mid s)$ reaches its minimum value, $\Phi^{MIP}(C \mid s)$, where MIP is the minimum information partition, and Φ^{MIP} stands for *integrated conceptual information*. Thus, if $\Phi^{MIP}(C \mid s) > 0$, no partition can divide the system into non-interacting, mutually independent parts. Moreover, the greater the value of Φ^{MIP}, the more irreducible the constellation of concepts generated by a particular set of elements[18]. Finally, according to the exclusion postulate, out of many possible constellations of concepts generated by overlapping sets of elements only one exists: the one that is maximally irreducible. Thus, one needs to evaluate Φ^{MIP} for all sets of elements s, i.e. s = A, B, C, AB, AC, BC, ABC[19]. The set of elements generating the constellation with the maximum value of Φ^{MIP} ($_{max}\Phi^{MIP}$, or *maximally integrated conceptual information*) constitutes the main *complex* within the overall system; the corresponding concept space (simplex) is called *qualia space*, and the constellation of concepts it generates – the *maximally integrated conceptual (information) structure* – is called a *quale Q*[20].

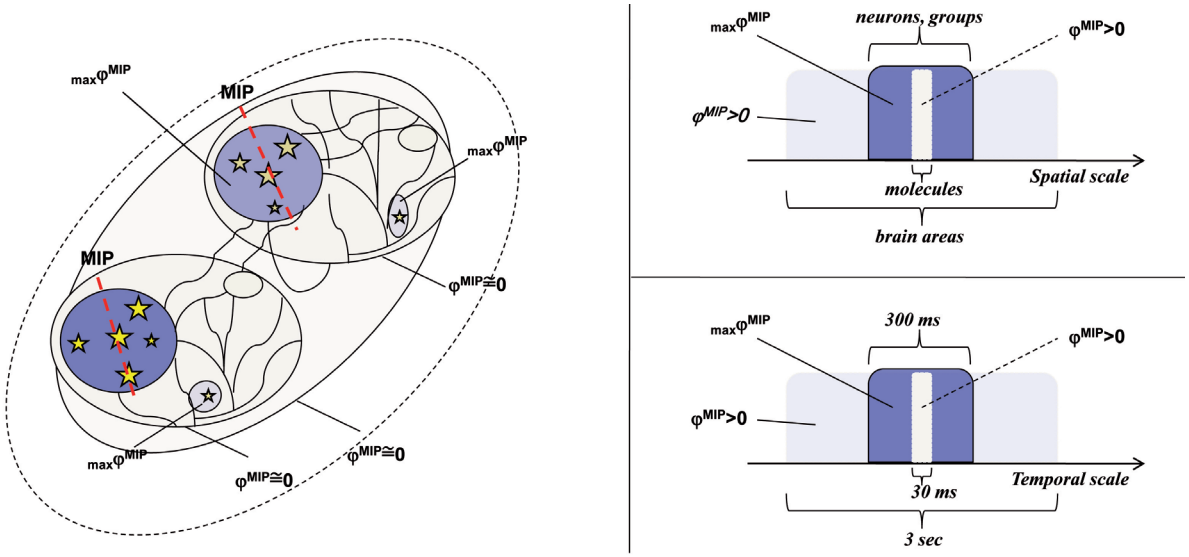
For example, an exhaustive analysis of the system in Fig. 4 shows that the full set ABC constitutes a complex, as no other set of elements yields integrated conceptual structures having a higher value of Φ^{MIP}. In larger systems, one would first identify the main complex and then, recursively, identify other complexes among the remaining elements. Therefore, a *complex* can be defined as *a set of elements generating a maximally irreducible constellation of concepts* (a maximally integrated conceptual structure). In essence, then, just like a concept specifies a particular, maximally integrated distribution of system states out of possible distributions (a point in concept space), a complex specifies a particular, maximally integrated conceptual structure (constellation of points) out of possible conceptual structures in concept space. As indicated by the information axiom, that constellation differs in its particular way from other possible constellations.

A schematic representation of a reduction of a system into complexes plus the residual interactions among them is illustrated in Fig. 5a. Note, for example, that due to the exclusion postulate, although complexes can interact, they cannot overlap. Thus, when two complexes of high $_{max}\Phi^{MIP}$ interact weakly, their union does not constitute a third complex, even though its Φ^{MIP} value may be > 0: once again, there is no need to postulate additional entities, because they would make no further difference beyond what is accounted by the two complexes of high $_{max}\Phi^{MIP}$ plus their weak interactions[21]. This is a direct application of Occam's razor: "entities should not be multiplied beyond necessity"[22]. We recognize this principle intuitively when we talk to each other: most people would assume that there are just two consciousness (complexes of $_{max}\Phi^{MIP}$) that interact a little, and not also a third consciousness (complex of lower Φ^{MIP}) that includes both speakers. In summary, a complex is an individual, informationally integrated *entity* that is maximally irreducible: i) it cannot be partitioned into more integrated parts; ii) it is not part of a more integrated system; iii) it is separated through a boundary from everything external to it (it *excludes* it). In this view, any system of elements 'condenses' into distinct, non-overlapping complexes that constitute local maxima of integrated conceptual information.

Optimal spatio-temporal grain

The exclusion postulate should be applied not only over sets of elements, but over different spatial and temporal scales. For any given system, one can group and average the states of several microelements into states of a smaller number of macro-elements. Similarly, one can group and average states over several micro-intervals into longer macro-intervals. For each spatio-temporal grain, one calculates CER, concepts (maximally irreducible CER), and complexes (sets of elements generating maximally integrated conceptual structures). By the exclusion postulate, a particular set of elements, over a particular spatio-temporal grain, will yield the max value of Φ^{MIP}, thereby excluding any overlapping subsets and spatio-temporal grains.

As an example, consider the brain: over which elements should one consider perturbations and the repertoire of possible states? A natural choice would be neurons, but other choices, such as neuronal groups at a coarser scale, or synapses at a finer scale, might also be considered, not to mention molecules and



- Integrated conceptual information Φ^{MIP} is the difference between the constellation (conceptual information structure) generated by the whole and that generated by its minimum information partition
- Maximally integrated conceptual information $_{max}\Phi^{MIP}$ is a local maximum of $_{max}\Phi^{MIP}$ over all sets of elements/space/time
- The corresponding maximally integrated conceptual information structure is a constellation or "shape" Q in qualia space

Fig. 5. - Complexes: maxima of integrated conceptual information over elements, space, and time. In the left panel, the blue ovals represent several separate complexes, i.e. local maxima of $_{max}\Phi^{MIP}$, each containing a schematic constellation, i.e. an integrated information structure comprising different concepts (stars). Each large blue oval – a main complex corresponding to an individual consciousness generated by a subset of neurons within the brain – is contained within a larger white oval that stands e.g. for the body, a system that does not constitute a complex and is thus not conscious. Inside the body, besides the main complex, are smaller complexes having very low $_{max}\Phi^{MIP}$ (only one shown) and presumably many smaller ones that are not represented. The curved lines represent interactions among parts of the body that remain outside individual complexes and thus outside consciousness. The large oval that encompasses both bodies indicates that the two consciousnesses interact within a larger system that is again not a complex and is thus not conscious. The outer dashed oval stands for the immediate environment. The right panels indicate that, within a system such as the brain, $_{max}\Phi^{MIP}$ will reach a maximum not only within a particular subset of elements but also at a particular spatio-temporal scale. See text for further explanation.

atoms. Importantly, under certain circumstances, a coarser spatial scale ('*macro*'-level) may produce a complex with higher values of Φ^{MIP} than a finer scale ('*micro*'-level), despite the smaller number of macro- compared to micro-elements. In principle, then, it should be possible to establish if in the brain consciousness is generated by neurons or groups of neurons. In this case the exclusion postulate would also mandate that the spatial scale at which Φ^{MIP} is maximal, be it neurons or neuronal groups, excludes finer or coarser groupings: there cannot be any superposition of (conscious) entities at different spatio-temporal scales if they share informational/causal interactions (Fig. 5b)[23].

Similar considerations apply to time. Integrated information can be measured at many temporal scales. Neurons can choose to spike or not at a scale of just a few milliseconds. However, consciousness appears to flow at a longer time scale, from tens of milliseconds to 2-3 seconds, usually reaching maximum vividness and distinctness at a few hundred milliseconds (Fig. 5c). IIT predicts that, despite the larger number of neural 'micro'-states (spikes/no spikes, every few milliseconds), Φ^{MIP} will be higher at the level of neural 'macro'-states (burst of spikes/no bursts, averaged over hundreds of milliseconds). This is likely the case because a set of neurons widely distributed over the cerebral cortex can interact cooperatively only if there is enough time to set up transiently stable firing patterns (attractors, see below) by allowing spikes to percolate forward and backward. Again, the exclusion postulate would mandate that, whatever the temporal scale that maximizes Φ^{MIP}, be it spikes or bursts, there cannot be

any superposition of (conscious) entities evolving at different temporal scales if they share informational/causal interactions[24 25].

Identity between maximally integrated conceptual structures (qualia) and experiences

In summary, a particular set of elements at a particular spatio-temporal scale yielding a maximum of integrated conceptual information ($_{max}\Phi^{MIP}$) constitutes a complex, a 'locus' of consciousness. The set of its concepts – maximally irreducible cause-effect repertoires ($_{max}\varphi^{MIP}>0$) specified by various subsets of elements within the complex – constitute a *maximally integrated conceptual information structure* or *quale* (Fig. 4) – a shape or constellation of points in qualia (concept) space[26].

Having defined complexes and qualia, IIT posits *identities* between phenomenological and informational/causal aspects of systems. The central identity is the following: *an experience is a maximally integrated conceptual (information) structure* or *quale* – that is, a maximally irreducible constellation of points in qualia space. Tentative corollaries of this identity include the following: i) the particular 'content' or quality of the experience is the shape of the maximally integrated conceptual structure in qualia space (the constellation of concepts); ii) a phenomenological distinction is a maximally irreducible cause-effect distinction (a concept). In other words, unless there is a mechanism that can generate a maximally irreducible cause-effect repertoire (concept) – a distinct point in the quale – there is no corresponding distinction in the experience the subject is having; iii) the intensity of each concept is its $_{max}\varphi^{MIP}$ value; iv) the 'richness' of an experience is the number of dimensions of the shape; v) the scope of the experience is the portion of qualia space spanned by its concepts; vi) the level of consciousness is the value of maximally integrated conceptual information $_{max}\Phi^{MIP}$; vii) the similarity between concepts is their distance in qualia space, given the appropriate metric; viii) clusters of nearby concepts form modalities and submodalities of experience; ix) the similarity between experiences would be given by the similarity between the corresponding shapes (see also the final section and Tononi, 2008, 2010), and so on.

In principle, then, given the "wiring diagram" and present state of a given system, IIT offers a way of specifying the maximally integrated conceptual structure it generates (if any)[27]. According to IIT, that structure completely specifies "what it is like to be" that particular mechanism in that particular state, whether that is a set of three interconnected logical gates in an OFF state; a complex of neurons within the brain of a bat spotting a fly through its sonar; or a complex of neurons within the brain of a human wondering about free will. In the latter examples, the full integrated conceptual structure is going to be extraordinarily complex and practically out of reach: we are not remotely close to having the full wiring diagram of the relevant portions of a rodent or human brain; even if we did, obtaining the precise quale would be computationally unfeasible[28]. Nevertheless, by comparing some overall features of the shapes of qualia generated by different systems or by the same system in different states, it should be possible to evaluate broad similarities and differences between experiences. IIT also implies that, if a collection of mechanisms does not give rise to a single maximally integrated conceptual structure, but to separate qualia each reaching a maximum of integrated conceptual information, then there is nothing it is like to be that collection, whether it is an array of electronic circuits, a heap of sand, a swarm of bats, or a crowd of humans.

Matching

So far, the maximally integrated conceptual structures generated by a system of elements have been considered in isolation from the environment – as is the case for the brain when it dreams. But of course it is also essential to consider how integrated conceptual structures are affected by the external world, especially since the mechanisms generating them become what they are through a long evolutionary history, developmental changes, and plastic changes due to interactions with the environment.

In any situation, a complex of high $_{max}\Phi^{MIP}$ has at its disposal a large number of concepts – maximally irreducible cause-effect repertoires specified within a single conceptual structure. These concepts allow the complex to understand the situation and act in it in a context-dependent, valuable fashion. It would be helpful to have a measure that assesses how well the integrated conceptual structure generated by an

adapted complex fits the causal structure of the environment. One way to do so is to define *cause-effect matching* (M) between a system and its environment as the difference between two terms, called *Capture* and *Mismatch*:

$$Matching = Capture - Mismatch$$

Capture is the minimum average difference $<D>$ between the constellations C when a complex interacts with its environment (C *World*), compared to when it is exposed to an uncorrelated, structureless environment (C *Noise*).

$$Capture = \min < D\ [\ C \mid s\ World, C \mid s\ Noise\ ] >$$

As before, *D* specifies a distance metric. *Capture* is an indication of how well the system *samples* the statistical structure of its environment (deviations from independence). Thus, high capture means that the system is highly sensitive to the correlations in the environment. The system can do so in two ways: on the input side, by sampling as many correlations as possible from the environment through a large sensory bandwidth and distributing these correlations efficiently within the brain through a specialized connectivity (thereby reflecting to what extent *World* deviates from *Noise,* Tononi et al., 1996). On the output side, an organism can extract more information by actively exploring its environment or modifying it to better pick up correlations, aided by a rich behavioral repertoire (Tononi et al., 1999). Note that the minimum is taken because to match system constellations generated with *World* and with *Noise* one should pair them in such a ways as to minimize the overall difference.

Mismatch is the minimum average difference $<D>$ between the constellations C when a complex interacts with its environment (C *World*), compared to when it is dreaming (C *Dream*), that is, when it is disconnected from the environment both on the input and the output sides.

$$Mismatch = \min < D\ [\ C \mid s\ World, C \mid s\ Dream\ ] >$$

Mismatch is an indication of how well the system *models* the statistics of its environment. Thus, low mismatch means that the system's causal information structure generates a good intrinsic model of its input. Again, the system can do so in two ways: by modifying its own connections so they generate a correlation structure similar to that induced by the environment (the system's *Dream* becomes a model of *World*). In this way 'memories' formed over a long time can help to disambiguate / fill in current inputs and, more generally, to predict many aspects of the environment (Tononi and Edelman, 1997). Another way is to change the environment by exploring it or modifying it to make inputs match its own values and expectations (*World* is made to conform to the system's '*Dream*'). In general, the interactions with the environment would have to match specific cause repertoires with specific effect repertoires in a way that yields perception-action cycles of high adaptive value: in short, the 'right' cause should lead to the 'right' effect

Note that the balance between the two terms in the expression for matching has two useful consequences: maximizing *Capture* ensures that the system does not minimize *Mismatch* simply by disconnecting from *World*. Conversely, minimizing *Mismatch* ensures that the system does not maximize *Capture* simply by becoming sensitive to the correlations in its input from *World* without developing a good generative model.

Importantly, since within a given system it is likely that similar states yield similar constellations, a simpler expression for matching can be obtained by considering differences between the probability distribution of system states **S**, rather than differences between sets of constellations C:

$$M = D\ [\mathbf{S}\ World, \mathbf{S}\ Noise] - D\ [\mathbf{S}\ World, \mathbf{S}\ Dream]$$

(note that, while the above expression is based on the distribution of system states, in principle the notion of matching can also be applied to the distribution of sequences of system states).

In the course of evolution, development, and learning, one would expect that the mechanisms of a system change in such a way as to increase matching. Capture should increase because, everything else being equal, an organism that obtains more information about the structure of the environment

is better off than one that obtains less information[29]. By contrast, mismatch should decrease since, everything else being equal, an organism having an internal generative model that matches well the overall causal structure of the environment is better off than one that is at the mercy of what happens here and now. Moreover, since high matching requires a large difference between cause-effect repertoires in the C-*World* and C-*Noise* conditions (capture), to optimize M a system should have *many different concepts*, i.e. have high $_{max}\Phi^{MIP}$. Put differently, large integrated conceptual structures, if well matched to the environment, provide a broad context to understand a situation and to plan an appropriate action[30]. If high M requires high $<_{max}\Phi^{MIP}>$, it follows that an increase in matching will tend to be associated with an increase in information integration and thus with an increase in consciousness. Finally, one would expect that the growth of matching and the associated growth of consciousness would also be a function of the complexity of the causal structure of the environment itself. In environments where survival can be achieved trusting on the efficient execution of nearly independent functions each with a narrow domain, context-dependency would not play a large role, and an organism would not need to achieve high values of M and $<_{max}\Phi^{MIP}>$. Conversely, rich environments that put a premium on context-sensitivity and memory, such as competitive social situations, should favor the evolution of organisms having high values of M and $<_{max}\Phi^{MIP}>$ (Albantakis et al., in preparation)[31].

Further theoretical considerations

The framework presented above will certainly need to be expanded and refined. However, even in its current from, it can shed some light on some broad theoretical issues that assume critical relevance if one takes integrated information to be a fundamental, intrinsic feature of reality (Tononi, 2008)[32]. One of these concerns the relationship between information and causation, another the potential advantages of systems with high capacity for information integration.

Information and causation

IIT assumes that mechanisms in a given state are *intrinsically* associated with certain maximally integrated conceptual structures, which they specify irrespective of external observers. Each concept exists if and only if an underlying 'causal' mechanisms is in working order and can 'choose' among alternatives, that is, select particular cause-effect pairs from past to future states that are compatible with its present state. Moreover, a concept exists if and only if it is maximally irreducible to subconcepts. Finally, an integrated conceptual structure itself only exists if it constitutes a maximum of integrated conceptual information over elements, space, and time.

From these premises, it is worth considering more closely the relationship between information and causation. Causation has often been interpreted as a correlation between successively observed events, as pointed out by David Hume: by observing that event 1 is reliably followed by event 8, we infer that 1 causes 8. This view of causation as strength (*reliability*) of a correlation is akin to the traditional view of information from the extrinsic perspective, as in Shannon's formulation, where a correlation between 1 and 8 means that one event carries information about the other (mutual information). Some more recent formulations, such as transfer entropy, impose the additional criterion of the directionality of prediction. However, it would seem that, to assess causation, it is not enough to observe a system, but it is necessary to perturb it and see what happens. In this vein, Judea Pearl has developed an interventional or *perturbation-based model of causation*: for instance, one does not merely observe the sequence 1,8, but one imposes input state 1 and sees whether event 8 is reliably observed (while the opposite may not be true). In this case one can conclude that 1 caused 8, going beyond a mere correlation.

Conceptualizing causation properly also requires the consideration of *counterfactuals*, that is, what would have happened if instead of event 1 some other event had occurred. For instance, would effect 8 still have happened if, instead of imposing 1, one had imposed perturbation 2,3,4,5, and so on? If it turned out that the system always ends up in state 8, we would begin to think that 3 was not so much caused by the preceding state 1, but rather, that 8 was inevitable. In other words, it would seem that, the less a cause is selective, the less of a cause it is. Some further thought indicates that properly considering counterfactuals ties the notion of causation even more closely with that of information, precisely because it implies *selectivity*. In the general case, it would

seem that one should consider all possible counterfactuals. That is, one should perturb the system in all possible ways and see what this does. This is exactly what is done by measuring cause-effect information (CEI) as defined above. CEI certainly depends on the *reliability* of the effects of a particular perturbation, as it decreases with noise. Cause-effect information also depends critically on *selectivity*: it is high if only some out of many past inputs can give rise to the present state of a mechanism and if this in turn can give rise to only some of many possible outputs[33 34 35]. It thus becomes apparent that the notion of causation, properly considered, is very similar to the intuitive idea of "difference that makes a difference", which is precisely what is captured by measuring cause-effect information.

If cause-effect information can indeed capture that causation must be related to differences that make a difference, what is the relation between causation and integrated information – the extent to which cause-effect information is irreducible, as established from the intrinsic perspective of a system? As was argued above, if a candidate cause-effect repertoire, as measured by CEI > 0, can be reduced to the product of independent components, as indicated by $\varphi^{MIP} = 0$, then there is no reason to posit its existence as an additional mechanism, because there are no further cause-effects to be accounted for beyond those accounted for by component mechanisms. That is, true causation requires not only that CEI > 0, but also that $\varphi^{MIP} > 0$. In other words, if what might at first look like a genuine a cause-effect relationship can be completely reduced to independent components, it makes no further difference, and thus it has no causal power.

An even stricter notion of cause is imposed by considering the notion of maximal integrated information ($_{max}\varphi^{MIP}$). As was also argued above, once an element is in a certain present state (say ON), from its intrinsic perspective it makes no difference which of the possible causes of its being ON may have occurred, so one can simply consider the maximally irreducible set of past causes and future effects (MICE) – those that make most of a difference. Based on the same notions one further identifies a set of elements – a complex – that specifies a maximally integrated information structure at an optimal spatio-temporal scale – one having $_{max}\Phi^{MIP}$. If such a local maximum of integrated information is indeed identical with consciousness, as claimed by IIT, it follows that a set of mechanisms in a state capable of generating consciousness also constitutes a *local maximum of causal power*. That is, consciousness itself is supremely causal[36].

The equivalence between consciousness and a local maximum of information/causation also suggests that, since consciousness is exclusive (there is only one consciousness with certain contents included, others excluded, and flowing at a particular spatio-temporal scale, given certain boundary conditions, causation itself may be exclusive: there is only one cause of any effect, the "core cause", and causation within "causal complexes" of elements only occurs at one spatio-temporal scale – the one that is maximally causal. Exclusion applied to causation has the virtue of immediately resolving the paradoxes posed by multiple causation. For example, when searching for the cause of an event (say, the pulling of a trigger), the cause should be found in the subset of elements that give rise to a maximum of integrated conceptual information with respect to the event (my conscious decision to pull the trigger). Any lesser cause (one that is less irreducible), including micro-level causes (the molecules in my brain), or proximal causes (the muscles in the finger) are excluded. That is, one should not double-count causes, just like one should not double-count information (causes should not be multiplied beyond necessity)[37].

Analyzing systems in terms of maxima of information/causation over many spatio-temporal scales also helps to address issues related to the possibility of causal emergence. For example, it is generally assumed that in physical systems causation happens at the micro-level, that a macro-level *supervenes* upon the micro-level (it is fixed once the micro-level is fixed), and that therefore a macro-level cannot exert any further causal power. On the other hand, if it can be shown that information/causation reaches a maximum at a macro- rather than at a micro-level (see note 22), then by the exclusion postulate there is true causal emergence: the macro-level *supersedes* the micro-level and excludes it from causation[38].

Concepts, questions, and qualia: Potential advantages of a maximally integrated conceptual structure over strictly modular structures

IIT suggests that a complex capable of generating a maximally integrated conceptual structure (a quale) should have some advantages over a collection of

independent modules. To gain some perspective on this issue, consider the system in Fig. 6, top. For simplicity, consider just 3 units (A,B,C) within a 'visual' area V1. Then there are units in area V2 that sample neighboring units in V1 and perform logical operations on the inputs (e.g. AND, XOR). In area V3 there are units that sample units of V2 and have a larger receptive field, which may extend to the entire visual field, and which also perform various logical operations. We assume that the units are linked not only by forward connections, but also by back-connections that implement further logical operations. In addition, units are linked by various mechanisms of competitive inhibition. Finally, the system is organized at the micro-level in such a way that after reentrant interactions between the units in different areas, it can settle into a limited number of metastable 'attractor' states over a macro time scale (see section on attractor dynamics below), and at this time scale it constitutes a complex.

This extreme caricature of the organization of the visual system can nevertheless be helpful in thinking about the kind of quale such a system might possibly generate. Let us consider the portion of qualia space specified over the three units A,B,C within V1 by the rest of the complex, focusing for simplicity on cause repertoires only (Fig. 6, bottom). For example, an AND unit V2-A that receives from V1-A and V1-B specifies, if it fires, that sub-states having the contiguous units V1-A and V1-B ON are compatible with its current state, and rules out all those states in which V1-A and V1-B were not both ON. Another unit in V2, the AND unit V2-B, specifies, if it is silent, that V1-B and V1-C could not possibly be both ON, and so on for other units in V2. A unit in V3, unit V3-A, which implements an XOR function of units in V2, may instead specify, if it is ON, that one and only one of the AND units in V2 must have been ON. Its firing then specifies a probability distribution over V1 that is compatible with any two contiguous units in V1 being ON together, but rules out many other firing patterns on V1. We can call this unit a position-invariant 'segment' unit. Similarly, a different logical function may yield a

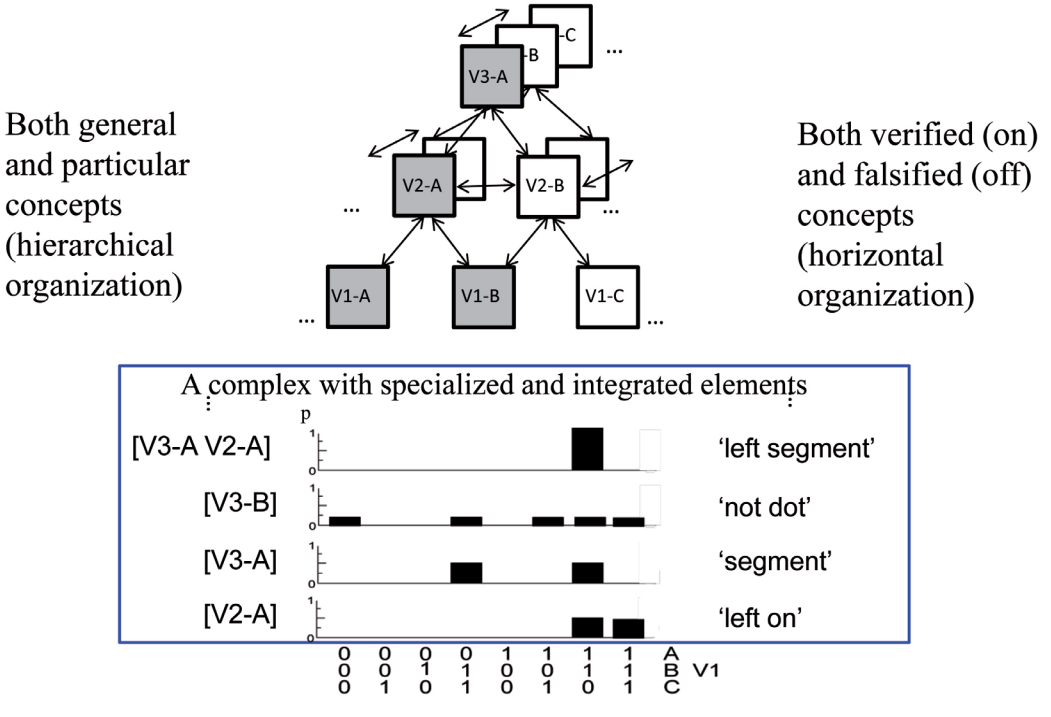


Fig. 6. - Schematic diagram of some of the concepts generated in a 'toy' visual system. Units that are ON are indicated in grey. Note that only some repertoires in the quale are shown (corresponding to some of the concepts discussed in the text), projected over the sub-states of just 3 units in V1. See text for explanation.

position invariant 'dot' unit in V3, which fires if there is an isolated unit ON in V1. And there can be a 'line' unit, which fires if three or more units in V1 are ON.

Let us consider just some of the cause repertoires over input units A,B,C in V1, ignoring the effect repertoires (Fig. 6, bottom). Concepts specified by V3 units are integrated over all 3 V1 units and, moreover, happen to correspond to invariants, such as dots, segments, and lines. By contrast, units in V2 specify concepts that are more restricted as well as non-invariant, such as an AND of nearby locations. Furthermore, concepts specified on V1 can be both 'positive' and 'negative'. For example, a 'segment' unit firing specifies a probability distribution over V1 that is compatible with there being a 'segment'; on the other hand a 'dot' unit that is silent also specifies a distribution, which corresponds to the negative concept 'not dot' and is less sharp. Also, concepts can be specified by combinations (subsets) of units taken together, as long as what is specified cannot be reduced to simpler concepts. Such irreducible concepts can be organized 'vertically' or 'horizontally'. As an example of 'vertical' organization, the unit in V3 signaling that there is a segment, together with a unit in V2 signaling which particular neighboring pixels in V1 are ON, together specify that there is a 'left segment'. Note that the V3 unit 'knows' that there must have been a segment, which is an important generalization, but does not know which particular one. By contrast, the V2 unit knows that its two inputs (on the left) are both ON, but it has no concept of a segment. Together, the V3 and V2 units know both the general concept (segment) and its particular location (left) – a more selective and integrated concept: integrated, because it is predicated on the entire input array, and maximally selective because it rules out all possible states except for V1-A = ON, V1-B = ON, V1-C = OFF. As an example of 'horizontal' organization, the segment unit in V3 knows that there is a segment, but it does not know that it is not a dot or a line, while another unit in V3 may know that there is no dot, but does not know about segments and lines. However, the segment, dot and line units together specify that there is a segment but not a dot or a line.

As is evident from this simple example, due to compositionality the number of possible concepts grows very rapidly with the number of elements in a complex[39], even if one considers just the repertoires specified over a subset of elements, such as those receiving input from the environment.

Clearly, only a fraction of all possible concepts is likely to be useful in a given environment. Useful concepts will be those that are most informative about the environment because they match its statistical regularities over space and time (see above). Nevertheless, in a rich environment, the number of concepts that would be relevant is presumably not small. Most importantly, it is likely that concepts relevant to a given environment are related in various ways – for instance, evidence indicating that a particular input is a segment also implies that it will not be a dot or a line; similarly if there is a segment, it will be either left or right. In these respects, it would seem that an integrated system (a complex), compared to a set of independent modules, may have some substantial advantages. Let us then contrast a complex, where many relevant concepts are implemented within a single, integrated conceptual structure (the quale), within a set of modules, where each relevant concept is implemented separately.

Economy of units/wiring and compositionality

A properly built complex, which can specify many different concepts using different combinations of the same units and connections, should be more economical than a collection of independent modules, where many (low-level) mechanisms would have to be duplicated. Within a complex, reciprocal (horizontal) inhibition between concepts helps to specify at once what an input is and what it is not. Also, high- and low-level concepts can be naturally combined (vertically) into hierarchical concepts ('there is a segment', 'there are two contiguous pixels ON at the left', can be combined to signal 'a segment on the left'). Moreover, an integrated architecture makes it possible to specify a large number of concepts not only over the input units, but over any combination of internal units, which can be highly abstract, related to arbitrary combinations of memories, or purely imaginary and divorced from sensory evidence. Note that, while it is conceivable to construct a system having only first order concepts, for example, in the extreme case, a neural network having 2^n hidden units each turning ON for a different state out of the 2^n states of n input units, and turning ON in turn a specific subset of output

units, such a system would be immensely expensive in terms of units and connections, it would have problems with adapting to a new environment, it would be extremely sensitive to noise, and so on. By contrast, a system exploiting *compositionality* (the powerset of 2^n combinations of n units) would need just n hidden units to represent the same number of concepts over the input units, would save on connections, simplify learning, be more resistant to noise, and so on.

Relational architecture

In a quale, concepts can be thought of as arranged along the lattice given by the power-set of subsets of elements in the complex. Individual units specify elementary concepts. Then, in the *context* of the concept specified by a particular unit (say unit V3-A), another unit (say V3-B), further specifies the repertoire, sharpening the concept. In the context of the concept specified by units V3-A and V3-B, unit V2-A further specifies the repertoire and sharpens the concept even more, and so on. Each incremental specification over an increasingly richer context is captured by additional points in the quale, capturing the compositionality and nesting of concepts[40]. In a strictly modular system, the relational structure of concepts is not inherent in how the various concepts are organized. The segment detection module does not know that there is no dot, as it has no concept of dot. Nor does it know where the segment is. The module detecting the two left pixels ON does, but it does not know that it is a segment. Since there are no mechanisms to generate the relevant distributions from the interaction of different modules, one would need a dedicated module for each relevant repertoire, and there would be no way of keeping track of how they are related.

Questions and answers

Another way to see the difference between integrated and modular systems is to consider questions and answers. It is useful to think of a question posed to a system as having a context and a specific query. Consider the question: "Is the segment on the left?" The *context* of a question, often implicit, refers to the assumptions that are necessary to 'understand' what the specific query is about. In the example of Fig. 6, the context includes the probability distribution compatible with 'there is a segment', in one of several positions, which is specified by unit V3-A. The *specific query* refers to the explicit issue at hand, here "Is the segment on the left?" It corresponds to the sharpening of the probability distribution of the context by further specifying "Are the left two pixels ON?" Clearly, a rich quale with many concepts implies that, at least in principle, the complex has at its disposal the answer to a very large number of questions, each of which can be put in the appropriate context[41]. Again, strictly modular systems would need a separate module for every question in every context, which rapidly leads to a combinatorial explosion. Moreover, there is the issue of how questions can be routed to the appropriate modules and answers can be sampled from them.

Access

In this respect, the integrated conceptual structure of the quale can be useful for querying the appropriate subsets of units and sampling their state by exploiting the relational structure of the quale. For example, the question "Is the segment on the left?" say triggered through the auditory modality[42], would be routed to unit V3-A (context) along connections – especially back-connections – that happen to be activated by the particular firing pattern the complex is in. The enhanced activation of V3 would then flow further to V2-A through activated back-connections. Finally, the enhanced activation of V2-A would exert differential effects elsewhere in the complex, eventually finding its way to output units primed for 'yes' or 'no'[43].

A complex can be computationally very effective, since accessing specific subsets of units serving as sources or targets can be done through a roadmap of connections that is both far-reaching (every unit in a complex can be reached by any other unit) and specific (it is possible to selectively reach particular subsets of units). In other words, the very mechanisms that allow a complex to generate many concepts that represent questions and provide answers, and to understand a question in both its context and specific query, are what allows the complex to direct the question to *access* the appropriate repertoires cooperatively. None of this is possible in modular systems, as one would need a readout wired *ad hoc* to the particular conjunction of modules that contain the relevant information[44].

Learning

An integrated system would also be advantageous for learning. Briefly put, a hierarchical, integrated organization with feedback connections can help learning in stages. For example, early on, low-level stages can discover local features over a small search space, then help higher levels to extract invariants over a larger search space. Later on, high-level stages can help lower levels by priming them based on contextual information. Lateral interactions can serve a similar role. None of this cooperativity in learning is available to modular systems.

Substrate for selection

An integrated system embodying a large number of nested relationships offers an efficient substrate for selectional processes. Designing a system with multiple feedback loops and interactions is notoriously difficult from an engineering perspective, which explains why the standard engineering approach is to design largely independent modules, minimizing their coupling to avoid unintended consequences. On the other hand, natural selection, as well as processes of neural selection (Edelman 1987), have no such qualms. Selectional processes can test countless integrated systems without regard for the complexity of the interactions among the constituting elements and simply choose based on the results: systems that work are elaborated further, while systems that fail for whatever reason are discarded.

Economy of understanding/control and incompressibility

Lastly, and more speculatively, to the extent that a complex having high $_{\max}\Phi^{\mathrm{MIP}}$ matches the causal structure of its environment, it can be argued that it would capture and 'understand' as well as 'control' that structure in a parsimonious manner that is informationally concise. This is because, first, the combinatorial organization of concepts in the quale permits to build a large relational structure based on a relatively small number of primitives. Second, if the environment's causal structure is itself highly relational (compositional, nested, etc.), only a *model* that is itself highly relational can provide an *explanatory/predictive structure* that is informationally highly compressed (non-redundant). Since an evolved organism having high $_{\max}\Phi^{\mathrm{MIP}}$ (and thus consciousness) would be sensitive to a large context of causal relationships in the environment, it should be more flexible than an organism equipped with a set of informationally separated processors, each of which has limited scope and understanding of the situation it finds itself in, which should make consciousness adaptive.

In summary, the integrated conceptual structure of the quale offers an economical way of assembling a large repertoire of different concepts, each applicable in many different contexts; ensures that concepts are arranged according to their relational structure; that they can easily be accessed accordingly; that many different concepts produce different effects on the system; and it facilitates learning, selection, understanding, and control. Even in the simple example of the 'left segment' discussed above, an appropriately built complex would know at the same time that there is a segment, that it is on the left, that it is not a dot (any dot) or a line (any line), and this knowledge is organized relationally. Arguably, only if a system is equipped with an integrated, relational organization of concepts can it hope to answer many different questions, including arbitrary questions, in an intelligent manner (Koch and Tononi, 2008), and only then can it be said to truly *understand* what it is seeing. As postulated here, the larger the context of understanding, the higher the degree of consciousness.

Some empirical considerations: Accounting for neuroanatomy and neurophysiology

As much as a theory of consciousness should be self-consistent and have heuristic value, ultimately it must be consistent with empirical data. Unfortunately, due to the special problems posed by assessing consciousness based primarily on behavior, any particular piece of data can often be difficult to interpret. However, the empirical evidence becomes less ambiguous when considered together, and indeed an important objective for a theory is explanatory power: being able to account for seemingly disparate data in a unified and parsimonious manner. The most relevant findings concerning the necessary and sufficient conditions for consciousness come from neuroanatomy, neurophysiology, and neuropsychology (Tononi and Laureys, 2009).

As discussed in previous work (Tononi, 2004, 2008), several observations concerning the neural substrate of consciousness fall into place within the IIT framework. Among them are: i) the association of consciousness with parts of the corticothalamic system but not with the cerebellar system; ii) the fact that neural activity over afferent and efferent pathways, and within cortico-subcortico-cortical loops remains unconscious; iii) the finding that even within cortex, some areas, such as the dorsal stream, do not seem to contribute to experience; iv) the special role that seems to be played by back-connections and supragranular layers; v) the evidence that consciousness can be split by anatomical and functional disconnections; vi) the loss of consciousness in generalized seizures despite intense, hypersynchronous firing; vii) the fading of consciousness during certain phases of sleep and anesthesia despite continuing neuronal activity; viii) the findings indicating a breakdown of effective connectivity and/or the occurrence of stereotypic responses in vegetative patients. The next section adds some considerations on the relationship between the temporal grain size of information integration and attractor dynamics in the corticothalamic complex.

Attractor dynamics in the corticothalamic complex

The corticothalamic system is the part of the brain that, if severely damaged, causes a loss of consciousness. Within the corticothalamic system, however, the situation is less clear, with respect to both necessity and sufficiency for consciousness (Tononi and Laureys, 2009). Can the cortex sustain consciousness without the thalamus? Is posterior cortex necessary and sufficient for consciousness, and does prefrontal cortex contribute at all? Are medial cortical regions necessary, or maybe the default network, perhaps as a connectional hub for intercortical or cortico-thalamo-cortical interactions? Do primary sensory and motor areas contribute at all? Does the dorsal stream in posterior cortex only contribute to behavior, but not to experience? What is the relative contribution of different cortical layers? Are supragranular layers necessary and sufficient for experience? Or are projection neurons in layer V critical? Do both excitatory and inhibitory neurons contribute to consciousness, or perhaps just some particular subset of cortical neurons? Or are the relevant elements groups of neurons rather than individual neurons? Can consciousness be sustained by feedforward connections only? Or are back-connections essential? Does every spike count, or only mean firing rates over hundreds of milliseconds? Finally, does the mode of firing matter, given the fading of consciousness during early slow wave sleep? Some of these questions are being investigated empirically, although definitive answers are hard to obtain. From the perspective of IIT, one can formulate a tentative scenario that may help to form a tentative model of possible neural substrates of consciousness, with the caveat that such a scenario at this point is still largely speculative.

It has been recognized at least since Lashley and Hebb that the massive interconnectivity within and among cortical areas (and with thalamus) provides an ideal substrate for cooperative dynamics among distributed neurons, which Hebb called cell assemblies and others called coalitions (Crick and Koch, 2003). A plausible scenario for characterizing such dynamics is in terms of *transient attractors* (Friston, 1997, 2000; Rabinovich et al., 2006; Deco et al., 2009). Simply put, neurons in the corticothalamic system seem coupled in such a way as to ensure the rapid emergence of firing patterns that are distributed over wide regions of the cortex, where some neurons are strongly activated, and many more are deactivated. These firing patterns remain stable (hence attractors) over a time scale of tens/hundreds of milliseconds, but then rapidly dissolve (hence transient), to make room for another transient attractor. Indeed, some EEG and MEG studies suggest that cortical activity patterns show brief periods of stability linked by even shorter periods of instability (Lehmann et al., 2009; Musso et al., 2010; Van de Ville et al., 2010). An example of this attractor dynamics from an early model of large-scale cortical networks is shown in Fig. 7 (based on Sporns et al., 1991; Tononi et al., 1992a).

Transient attractors and integration

According to IIT, several aspects of the organization of the corticothalamic system and of transient attractor dynamics appear well suited to information integration. The corticothalamic system includes strong local links as well as a network of long-range connections among nearby and distant areas, many of them reciprocal, giving rise to reentrant loops that favor

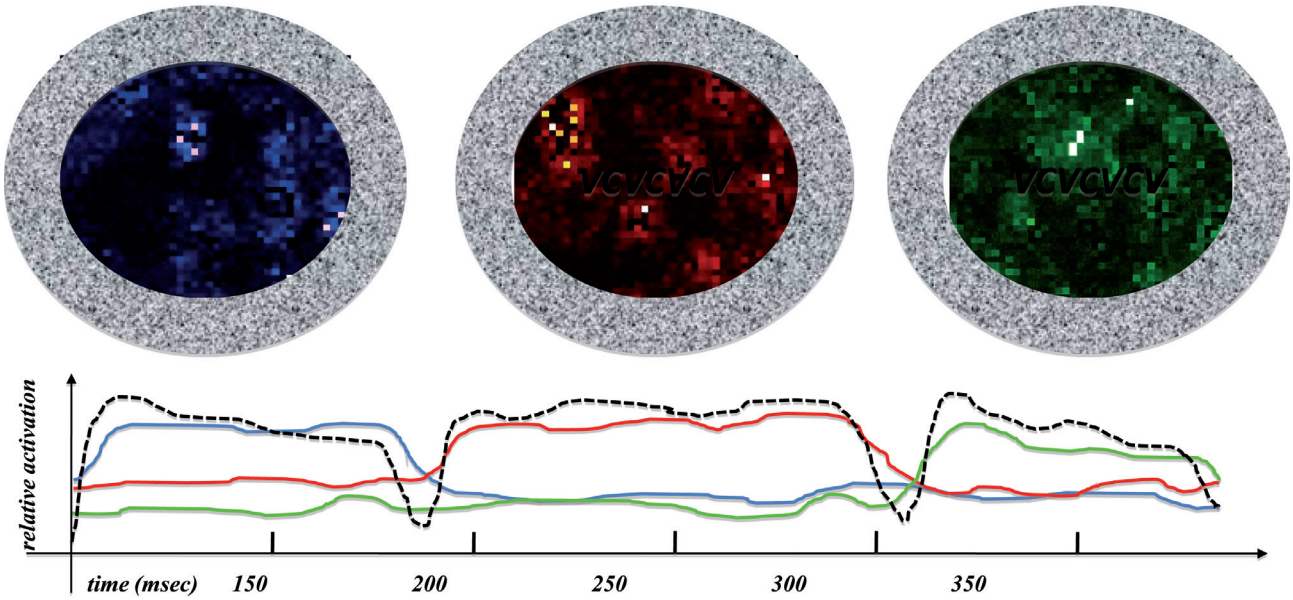


Fig. 7. - Top. Three snapshots of spontaneously generated transient attractors in a large-scale model of the cerebral cortex, based on (Sporns et al., 1991; Tononi et al., 1992a). Each colored core shows a meta-stable state (transient attractor) occurring over a central complex of units linked by specialized cortex-like connectivity. The firing patterns displayed were averaged over 50 milliseconds (200 time steps), with highly active units indicated in white and inactive ones in black. The grey scale 'halo' represents separate chains of units each independently connected to units of the central complex, from which they remain excluded. Bottom. Time course of the relative activation of each of the 3 transients attractors (color coded). Note the rapid ignition, relative stability, and fast quenching of each transient attractor, and the emergence of the next one. The dashed line shows the accompanying fluctuations of the short-term enhancement in the efficacy of activated synapses (averaged over all synapses), which boosts both the ignition and the quenching of the attractors.

integration (Sporns, 2010; Stratton and Wiles, 2010). Connectional hubs along the medial surface of the cortex may facilitate the interaction of distant cortical regions; more diffuse projections from thalamic matrix cells may provide a shared background of excitability that also facilitates long-range interactions; the reticular thalamic nucleus may provide strong inhibitory coupling among distributed cortical areas.

Several dynamic factors also help long range, effective interactions. The rapid formation of attractors may be boosted by short-term strengthening of activated synapses; their dissolution after a brief period of stability may be brought about by the ensuing short-term depression of synapses, as well as by destabilizing signals from neuromodulatory systems (Sporns et al., 1991; Tononi et al., 1992a). The time constants of neuronal integration, of various intrinsic currents, of AMPA and GABA receptors, and especially of NMDA receptors, ranging from tens to hundreds of milliseconds, also seem well suited to enhancing, sustaining, and terminating effective interactions in such a way that cooperative dynamics can extend to much of the cortex and yet remain flexible.

It is expected that the time scale of attractor formation and dissolution would correspond to the macro-time scale at which integrated information reaches a maximum (see above, spatio-temporal grain). If that temporal grain is indeed that of meta-stable transient attractors, then the *micro-level interactions* among neurons at the time scale of a few milliseconds, although they constitute the microstructure underlying *attractor macro-states*, would have no counterpart in phenomenology, consistent with the longer time scale at which consciousness seems to flow.

It is also important that, during conscious states, neurons are usually poised at the edge of firing, and are thus extremely reactive to perturbations, in line with work on avalanches, criticality, and neural 'noise' (Beggs 2008; Sporns 2010; Chen et al., 2010; Deco et al., 2011). Oscillatory dynamics may enhance long-range interactions, and synchronization across multiple frequency bands is both

an indication that short- and long-range effective interactions are taking place, as well as a mechanism to make such interactions more effective (Tononi et al.,1992a; Singer, 2009; Buzsáki, 2010).

IIT emphasizes that 'winning' (active) neurons are informationally meaningful only if considered in the context of their 'losing' (inactive) counterparts within the same complex. Thus, the excitatory connectivity must be complemented by an adequate inhibitory background to ensure that the network behaves as a single entity, both dynamically (attractor) and informationally/causally (complex). Short- and long-range excitatory connections in cortex innervate local inhibitory interneurons, which enforce a competitive dynamics that depends on the precise timing and balance of excitation/inhibition. It is also likely that the thalamic reticular nucleus plays a role in enforcing integrated activity patterns, by coupling local increases in activity with decreases elsewhere in the corticothalamic complex. Together, these various mechanisms ensure that attractor dynamics within the corticothalamic system is usually integrated within a single complex, meaning that it cannot be decomposed into the independent dynamics of separate attractors.

Transient attractors and information

The corticothalamic system is not only remarkably integrated, but it is just as remarkably specialized, at multiple scales: different cortical lobes, areas, groups of neurons and even individual neurons are selectively activated by different input patterns. Functional specialization is essential to ensure that, within a single corticothalamic complex, there is a large repertoire of transient attractors, which is a requisite for having a complex of high $_{\max}\Phi^{\mathrm{MIP}}$. A large repertoire of distinguishable attractors means that, when the complex is in any particular (meta-stable) attractor state, it is highly informative about which previous state would have caused it and which subsequent state it may effect. While classic attractor networks (with symmetric, all-to-all connectivity) have a limited repertoire of inflexible attractor states (measured as storage capacity), several features of the corticothalamic system may greatly increase the repertoire and make it more easily accessible (*'fluid' attractors*). These features include sparse, asymmetric connectivity organized at both short- and long spatial ranges, various kinds of inhibitory mechanisms, sparse activity patterns (at any given time, only relatively few neurons are strongly active), and short term changes in synaptic and intrinsic conductances.

The asymmetry between converging forward connections and diverging back-connections seems especially important in ensuring the compositionality of information integration. Neurons higher up in the sensory hierarchy are selective for higher level concepts (see above) or invariants (a face, anywhere in the visual field), whereas neurons in the early stages of sensory hierarchies are selective for lower level concepts (a vertical edge in a particular portion of the visual field). Through converging forward, driving connections, low-level neurons can select which high-level neurons should fire. Conversely, through diverging, modulatory back-connections high-level neurons can reinforce the activation of low-level neurons and constitute stable firing 'cliques' that link general and particular concepts. Such cliques could underlie hierarchical attractors (Bělohlavek, 2000; Gros, 2009; Wennekers, 2009). An interesting possibility is that, in such hierarchically organized attractors, some portions of a transient attractor (the 'head' or 'pivot', perhaps localized more frontally) may be stable for longer intervals than other, nested portions (the 'body and limbs'), perhaps located closer to sensory areas (Braun and Mattia, 2010).

Finally, the abundance of "loops" of various length that connect each neuron to itself through the intermediary of different subsets of neurons and different subsets of output/input connections, suggest that attractors, in addition to being transient and hierarchically organized, may be implemented not just as stationary patterns of activity, but as *sequences* of activations though such loops.

The combination of functional specialization and integration, together with a hierarchical organization of mechanisms linking general and particular concepts, is well suited to the generation of qualia containing many points, yielding a conscious experience that is high in quantity (integrated information) and extremely specific in quality (a shape breaking many symmetries).

Spontaneous activity and responsiveness to environmental inputs

It has been known for a long time that the corticothalamic system is spontaneously active, even during sleep. Moreover, this ongoing activity, even in the absence of environmental inputs, can be associated

with consciousness, as illustrated vividly by dreams. In fact, dreams may offer the most direct demonstration of the intrinsic dynamics of transient attractors and of the informational structure of consciousness. For example, dreams suggest the hierarchical nature of such dynamics, with some portions of the attractor lasting for longer times (dream setting, narrative) and others switching more rapidly, giving rise to a changing kaleidoscope of scenes and objects. The particular sequence of attractors in a dream is probably biased by the priming of subsets of connections by the previous attractor, thereby revealing associative links. During wakefulness, instead, while spontaneous activity persists, the selection of transient attractors is biased towards those that best capture the flow of inputs from the environment, possibly aided by a reset generated by novelty signals. An interesting question is whether during wakefulness transient attractors are 'ignited' and shaped bottom-up, thanks to the driving effect of forward connections, whereas during dreaming the ignition and shaping are top-down (Nir and Tononi, 2010).

A geometrical agenda: Accounting for phenomenology

Besides accounting for experimental data, ultimately a theory of consciousness should also shed some light on the seemingly ineffable qualitative properties of phenomenology. For example, what is responsible for the particular temporal scale at which experience flows? What is responsible for the fact that much of experience appears to be organized in space? What accounts for the indisputable organization of experience into modalities and submodalities? What makes a color different from a smell? And what makes the color red feel the way it feels, and different from blue? Viewing an experience as a shape in qualia space implies that features of experience that seem impossible to account for in neural terms – like the redness of red or the differences between spatial vision and color vision or between vision and sound – should instead be accounted for in mathematical terms. As briefly suggested below and elsewhere (Tononi, 2004, 2008; Balduzzi and Tononi, 2009), one can envision a close association between phenomenology and the geometry of qualia. Some identities were already pointed out above:

i) The particular 'content' or quality of the experience is the shape of the maximally integrated conceptual structure in qualia space (its shape Q) generated by a complex.
ii) A phenomenological distinction is a maximally irreducible cause-effect distinction (a concept).
iii) The intensity of each concept is its $_{max}\varphi^{MIP}$ value.
iv) The 'richness' of an experience is the number of dimensions of the shape.
v) The scope of the experience is the portion of qualia space spanned by its concepts.
vi) The level of consciousness is $_{max}\Phi^{MIP}$ – the maximally integrated conceptual information generated by the complex.
vii) Similarities and dissimilarities between concepts should translate into closeness/distance in qualia space (blue is closer to red than it is to a sound). Similarities/dissimilarities between experiences should translate to objective measures of *similarity/dissimilarity between shapes* in qualia space (related to the number and kinds of symmetries involved in specifying shapes or operations needed to transform one shape into another).
viii) The classic sensory modalities and submodalities (sight, hearing, touch, smell, taste, and within sight color and shape) would correspond within a quale to subsets of points that are clustered together in qualia space (modes and submodes).
ix) A quale in the narrow sense (the blueness of blue) is a Q-fold: the sub-shape of Q that is lost when the contribution or a particular element is lost, either in the concept it specified alone, or in those it specified in combination with other elements.
x) Even elementary experiences (qualia in the narrow sense, such as pure blue) translate to highly complex shapes in qualia space and cannot be reduced to anything less. This is because all the mechanisms specifying blue as opposed to every perceivable color, a full-field spatial experience as opposed to a composite experience including shapes and movement, a primarily visual experience as opposed to an auditory or olfactory one) must be operational to specify how that particular experience differs, in its own specific way, from every other possible experience, and make it what it is. In other words, the mechanisms specifying the color blue operate within the context of many other mechanisms that act, informationally, within a single quale.
xi) Homogeneous experiences (a blue cloudless sky) would correspond to a homogeneous shape,

whereas composite experiences (a cluttered desk) would correspond to a composite shape with many distinguishable sub-shapes (modes and sub-modes).
xii) The compositional structure of many experiences would have an informational counterpart within the structure of the quale (Fig. 4). For example, seeing a square in the left lower field of vision implies the specifications, within the same quale, of many concepts, i.e. probability distributions. These include distributions specifying that the invariant 'square' is present (a repertoire specifying particular configurations of inputs compatible with the concept 'square', irrespective of its position in the visual field), as well as repertoires specifying its actual details (where each edge is and how it is oriented). Moreover, this 'vertical', 'hierarchical' compositionality must be complemented by the 'horizontal' specification of what a square is not, that is, by a large number of points specifying that alternative invariants are absent (repertoires specifying the concepts 'not a triangle', 'not a circle', 'not a face' and so on). Finally, there may be an 'associational' specification of concepts tied to a square (cube, dice, checkerboard etc.). Only if the quale contains all the relevant concepts and informational relationships, can one say that the complex 'understands' the square or, which is the same, that it 'sees' it consciously.
xiii) 'Categorically' structured experiences (taste, smell, color) and 'topographically' structured ones (visual space) would correspond to different subshapes in qualia space, such as *pyramid-like* and *grid-like* shapes, which emerge naturally from the underlying neuroanatomy.
xiv) The refinement of experience that occurs through learning (as when one becomes a connoisseur in some domain) would translate to a corresponding refinement of shapes in qualia space, due to splitting of concepts mediated by changes in the underlying connectivity.
xv) Unconscious determinants of experience (e.g. the automatic parsing of sound streams into audible words) would be 'hidden' because the underlying mechanisms are outside the complex or occur at other spatial and temporal scales. This can happen, for instance, through units that are shared between a main complex and smaller complexes that serve as input or output channels or through loops carrying out local computations, whose internal information structure remains isolated from that of the main complex[45].

Conclusion

In summary, IIT attempts to provide a principled approach for translating the seemingly ineffable qualitative properties of phenomenology into the language of mathematics. Ideally, when sufficiently developed, such language should permit the geometric characterization of phenomenological properties generated by the human brain as well as by other brains, natural or artificial. The theory also provides a parsimonious, self-consistent framework that attempts to account for key neuroanatomical, neurophysiological, and neuropsychological observations. To make progress, the theory will need extensive mathematical developments, practical ways of measuring integrated information (see for example Barrett and Seth, 2011; Griffith et al., in press; Oizumi et al., in press; van Veen et al., in press), complexes, and the integrated conceptual structures generated by large systems[46]. As one would expect with consciousness, this will require back and forth validation between theoretical models and empirical findings.

Acknowledgements

I thank Chiara Cirelli, Lice Ghilardi, Christof Koch, Barry van Veen, Chris Adami, Larissa Albantakis, Mèlanie Boly, Virgil Griffith, Atif Hashmi, Arend Hintze, Erik Hoel, Matteo Mainetti, Andy Nere, Masafumi Oizumi, Umberto Olcese, and Puneet Rana for many helpful discussions. I am especially grateful to V. Griffith for his contribution to characterizing the concept of synergy and its relation to integrated information; to M. Oizumi for his help in characterizing cause-effect repertoires; to M. Mainetti for his help in characterizing the proper metric for conceptual spaces, and to Erik Hoel, Larissa Albantakis, and Christof Koch for many discussion concerning the relationship between information and causation. For developing the software used to compute integrated conceptual structures I am indebted to M. Oizumi, B. Shababo, L. Albantakis, A. Nere, A. Hashmi, U. Olcese, and P. Rana. This work was supported by a Paul Allen Family Foundation grant and by the McDonnell Foundation.

Notes

1 The expression "difference that makes a difference" was coined by (Bateson, 1972).

2 Descartes started his philosophical investigations from the axiom '*cogito ergo sum*', though his 'cogito' emphasized the thinking aspect of consciousness rather than the more general notion of having an experience.

3 Contrast this *intrinsic* perspective – the one of the system itself – with the *extrinsic* perspective of an external observer: the observer can ask how information is encoded, communicated or stored given the system's state transitions, the observer's expectations (prior distribution, e.g. based on observing the system), and assumptions about the system. Also, note that the term information here is used in a way that is closer to its original Latin meaning of "giving form" than to its use in communication theory. Thus, the information postulate can be interpreted as saying that a mechanism in a state does something only if it "gives form" to its past and future, by causally *constraining* what otherwise would be utter disorder (maximum entropy). The integration postulate says that a mechanism in a state does something only if it does more than its parts, i.e. it is irreducible. The exclusion postulate says that a mechanism in a state does only one thing (an IF-THEN "transformation") – the one that is maximally irreducible, thus avoiding multiple causation.

4 This presentation is an update of previous expositions (Tononi, 2004, 2008; Balduzzi and Tononi, 2008, 2009; Tononi, 2010). For example, repertoires are defined on subsets of elements rather than subsets of connections, based on the consideration that elements have 'states' and can thus make statements (ON/OFF, YES/NO, TRUE/FALSE). Moreover, instead of just input repertoires, this updated version makes use of cause-effect repertoires, which pair what elements specify about past and future. Also, only repertoires having $\varphi^{MIP} > 0$ are considered as potentially existing, and of those only those having $_{max}\varphi^{MIP}$ as actually existing. Finally, systems are analyzed bottom-up rather than top-down: after designating a particular system of elements at a particular spatio-temporal scale (candidate system), one first finds all maximally irreducible cause-effect repertoires (concepts), starting from first-level concepts (generated by individual elements), then second-level concepts (generated by two-plets of elements), and based on the constellation constituted by all the concepts determine the irreducibility of the system. In previous work, one would first evaluate if a system was irreducible as a whole, and then one would identify its concepts.

5 Measures of causal information flow are been investigated in the context of complex and adaptive systems (see for example Ay and Polani, 2006).

6 An element is a unit that receives input connections, performs an operation on those inputs, and outputs its new state to its output connections (input/output functions defining each mechanism can be deterministic or probabilistic). Such input-output functions yield the transition probability matrix (TPM). It is assumed here that elements implement the most 'elementary' mechanisms – they are memoryless 'micro-elements' – e.g. logical gates or linear threshold units. In principle, one could use just a single kind of such micro-elements, such as a NOR gate, to construct any other logical gate. More complicated units, having memory and capable of internal processing, can be constructed out of micro-elements. However, to be considered as elements in their own right (as opposed to a collection of elementary mechanisms), such units would need to constitute macro-elements (generating more information than the constituting micro-elements, in space or time, see later sections), and their internal processing would not be communicable elsewhere in the system.

7 The difference D(p,q) between two probability distributions p and q can be measured in various ways. In previous work the Kullback-Leibler divergence $D_{KL}(p||q) = \sum p_i \log_2 p_i/q_i$ was used. D_{KL} has several useful properties but it is not symmetric and is not bounded, which makes it not ideal for evaluating divergences when q is not the maximum entropy distribution u (in that case, for discrete distributions, D_{KL} is the same as the difference in entropy between the distributions). Moreover, not being a metric, D_{KL} does not take into account whether some states of the system are closer than others (e.g. whether [0 0 0] is closer to [0 0 1] than to [1 1 1]). Distance measures in an appropriate metric space (see concept space below), such as the Wasserstein distance (also known as earth mover's distance), may be better suited at capturing differences between distributions from the intrinsic perspective of a system. More generally, one would want a measure of the difference made by a mechanism (e.g. before and after a partition) not just as reduction of uncertainty (information as communication capacity, in the Shannon sense), but as specification – what it takes to transform something (here a distribution) and make it into something else (information as giving a particular form, in the classic sense of the word). Perhaps the most general way to do so is to consider the *information distance* between two objects, i.e. the maximum of the algorithmic complexity of one object given the other (shortest program that computes one given the other), which may ultimately correspond to the amount of thermodynamic work required to transform one into the other in the most efficient manner (Bennett et al. 1998; see also the notion of logical depth, which takes into account how long the

program takes). However, algorithmic complexity is a non-computable lower bound, which means that one needs to resort to more practical approximations, such as appropriate compressors. Moreover, one could argue that the appropriate distance between two distributions should be constrained by the intrinsic nature of the space within which the distributions live. This would lead back to earth mover's distance, in which case the minimal program would be the one specifying how to optimally move the earth (probability values) from one distribution to the other.

8 Partitions, indicated by x, can be evaluated by performing the same computations after injecting noise ($do(H^{max})$) in the partitioned links in the input-output matrix. To fairly compare different partitions to find the MIP, it is necessary to normalize by the information capacity of each partition.

9 where the empty set [] is only allowed on either P or S, but not both.

10 If several CER(S) yield the same max, one takes the CER(S) of largest scope (accounting for the most), where $\varphi^{MIP}(S)>0$, its subsets R have lower or at most equal φ^{MIP}, and its supersets T have lower φ^{MIP}: $\varphi^{MIP}(R) \leq \varphi^{MIP}(S) > \varphi^{MIP}(T)$, for all $R \in S$ and all $T \in S$. If there are multiple maximal CER(S) each with the same scope, then at any given time only one is realized as a concept, although which one is indeterminate.

11 Finding the partition yielding $_{max}\varphi^{MIP}$ and corresponding to the maximally irreducible set of causes-effects (MICE) is conceptually related to max flow – min cut problems.

12 One could say that trying various CER and their partitions to find $_{max}\varphi^{MIP}$ is the informational/causal equivalent of "cutting to the chase". It is also related to finding the optimal tradeoff between the transmission of relevant information and the compression/efficiency of the channel, see for example (Creutzig et al., 2009).

13 In neural terms, the fact that, out of all possible causes of a neuron's firing, the input that actually caused its firing remains undecidable from the intrinsic perspective, also means that "illusions" are inevitable. Based on the exclusion postulate, the intrinsic perspective entails the simplifying attribution of cause always to the core (most irreducible) cause, rightly or wrongly. Usually, in an adapted system, the actual cause and the core cause will be similar enough, but occasionally the actual cause may be quite different from the core cause, in which case an "illusion" ensues (this also applies to the case of a neuron's firing being caused by subtle microstimulation).

14 The exclusion postulate is related to the principle of sufficient reason – in fact, it enforces a principle of *least* reducible reason; to the principle of least action; to maximum likelihood approaches and to information minimization/compression (though it is causal, not just statistical); and of course ultimately to Occam's razor.

15 In this example, the cause repertoire component of a concept (backward, input, retrodictive, receptive concept) can be taken to refer to a classic *invariant* – a set of inputs equivalently compatible with the present state of a certain mechanism (e.g. tables, faces, places, and so on); the effect repertoire component (forward, output, predictive, projective concept) can be taken to refer to 'Gibsonian' *affordances* – a set of outputs equivalently compatible with the present state of a certain mechanism (e.g. the consequences/associations/actions primed by seeing a table, face, place, and so on).

16 The *scope* or 'volume' within concept space can be interpreted as a measure of how extensively the concepts sample concept space. Ideally, they should be distributed in such a way that, together, they are as informative as possible about the concept simplex. This last point can be appreciated by comparing systems in which many different subsets specify similar concepts (nearby points) or, in the limit, identical concepts (a single point), with systems in which different subsets specify very different concepts (distant points). For the same number and intensity of concepts, the first kinds of systems generate mostly redundant information, covering only a small corner of the concept simplex (small scope) from a single 'purview' (sharing the same specialization) and have thus very limited 'understanding/control over the system's states. By contrast, the second kinds of systems generate information from many different purviews (different specializations) covering a much larger portion of the concept simplex (large scope), and have thus much greater understanding/control over the system's states. To evaluate the scope of the sampling of the concept simplex by a system's concepts, one can associate each concept with a ball of unit volume, where the unit volume is obtained by packing the simplex with the 2^n concepts (the maximum number of possible concepts for a system of n elements, corresponding to its powerset) generated by a system composed of n independent elements (having self-connections) that are arranged in a product structure. Compared to this product structure, a typical system will likely generate fewer concepts, and balls of unit volume centered around them may overlap, resulting in a less informative sampling (redundancy). The *conceptual scope* is thus the overall volume of the concept simplex (without counting overlaps) sampled by the actual concepts generated by a system, assuming unit volume for each. One can then weigh the scope by multiplying each unit volume by its intensity, i.e. the amount of integrated information generated by

that concept ($_{max}\varphi^{MIP}$). For this purpose, when there is overlap, only the volume with highest $_{max}\varphi^{MIP}$ enters the multiplication.

17 Note that constellations of concepts must satisfy several requirements: i) they must be physically realizable; ii) they must be self-consistent (that is, concepts that exclude/contradict each other cannot coexist; i.e. their product should never yield a distribution with zeros everywhere); iii) they must be irreducible. If these requirements are satisfied, ideally a constellation of concepts should also: i) have as many concepts as possible; ii) they should be as irreducible as possible; iii) they should be as informative as possible about concept space, i.e. sample it as uniformly as possible (acting as representative "prototypes" of possible contingencies).

18 As for concepts, evaluating the difference *D* between the two constellations (here, those of the unpartitioned and the partitioned system) can be done by considering a metric based on the earth mover's distance, where $_{max}\varphi^{MIP}$ values play the role of the weight to be moved. Unlike with probabilities, $_{max}\varphi^{MIP}$ values of concepts do not sum to a fixed value. It is thus necessary to assume that the $_{max}\varphi^{MIP}$ value of concepts lost due to a partition is moved to the maximum entropy distribution (Mainetti et al., in preparation). Alternatively, the difference between constellations could be measured by considering the information distance between them, i.e. the maximum of the algorithmic complexity of one given the other (Bennett et al. 1998). Since this lower bound is generally non-computable, one needs to resort to measures such as minimum description length or generalized Kernel distance (e.g. (Joshi et al. 2011)) on the cause and effect subspaces. Since the sum of $_{max}\varphi^{MIP}$ values can differ between two constellations, one needs to consider not only the difference between constellations in terms of the relative positions of their concepts, but also in terms of their $_{max}\varphi^{MIP}$. Moreover, the measure *D* should be sensitive to whether sets of concepts in each constellation occupy nearby positions in concept space or not: for example, it will take less to go from one constellation to the other when the starting/ending points are close to each other than when they are not. As a consequence, constellations of concepts generated by highly homogeneous systems, in which most concepts are identical or nearly so, will differ much less from their partitioned counterparts than functionally specialized systems, where concepts are widely apart, i.e. *specific* (larger scope, see above; in addition to being generally much more numerous and integrated).

19 After "freezing" he links between the set and its environment as well as any other internal parameter and treating them as "boundary conditions." Thus, after selecting a candidate system of elements at a particular spatio-temporal scale, among the selected elements one evaluates cause-effect information and irreducibility by imposing all possible states as counterfactuals and examining the consequences. All other variables and parameters internal to the system or at its interface with external elements or forces are considered as boundary conditions and treated as "factuals", i.e. they are taken as fixed in their actual state. Such boundary conditions can include inputs from external elements through which the system interacts with its environment, or factors that sustain the system's functioning, such as energy supply, neuromodulators that promote excitability, and so on, at various spatio-temporal scales.

20 Within an integrated conceptual structure, one can distinguish a backward portion (specified by the cause repertoires), or *understanding*; and a forward portion (specified by the effect repertoires), or *control*.

21 Unless, of course, the interactions become so strong that $_{max}\Phi^{MIP}$ for the union exceeds that of each part, in which case the parts merge into a single complex.

22 Occam's razor conventional formulation, "entia non sunt multiplicanda praeter necessitatem", is probably due not to Occam or his teacher Duns Scotus, but to John Ponce. It has important applications in the context of Solomonoff theory of inductive inference and compressibility (Solomonoff, 1964), see also (Hutter, 2005). If one can compress a wiring diagram into a product of smaller diagrams (e.g. by finding k-connected subgraphs) plus some residual terms, one identifies separate integrated conceptual information entities that cannot be reduced further (complexes), and beyond which no additional 'higher' entities exist. Each complex is then characterized by a particular integrated conceptual structure, within which different repertoires specified by subsets of elements exist only to the extent that they are not reducible.

23 If they do not, then a spatio-temporal scale can be contained within another without overlap, at most providing a "boundary condition" (fixed parameter) for the interactions at the other scale.

24 For a macro-level to beat a micro-level, despite the much larger number of states that are available to the micro-level, some features are especially important: i) the presence of some degree of indeterminacy at the micro-level (due to intrinsic noise or to perturbations from the environment); ii) many-to-one mapping, such that many input states can produce the same output state, giving rise to irreversibility; iii) macro-mechanisms structured in such a way that they group noisy micro-states together in an advantageous manner; iv) the fact that, from the intrinsic perspective of the macro-system, all possible perturbations

(i.e. counterfactuals) must be conceived as applied to macro-states. This means that the actual distribution of micro-states underlying the macro-level distribution will be different from their micro-level maximum entropy distribution, thus accounting for emergence without violating supervenience. In summary, the level at which 'things' really exist in and of themselves, i.e. from the intrinsic perspective, in both space and time, is the level at which $_{max}\Phi^{MIP}$ is maximized – that is, the level at which 'causal power' is maximal. In other words, what really exists (and excludes any other level) is what makes the most difference – and this level is not necessarily the micro-level as is often assumed in reductionist accounts.

25 An important issue raised by the micro-macro distinction concerns computer simulations of, for example, conscious brains. If the logic gates that ultimately are responsible for simulating the informational/causal interactions among neurons that generate consciousness within the brain, cannot themselves be macroed into elements and intervals corresponding, say, to neurons over hundred milliseconds, then the intrinsic informational/casual properties of those macro-elements/intervals (the corresponding cause-effect repertoires etc.) would not exist intrinsically within the computer, yielding a true "zombie." Except in the light of a theory validated with other means, how could one tell the difference?

26 The terms anatomical, functional and effective connectivity are commonly used, although sometimes with different interpretations, especially for effective connectivity. In a general sense, one should distinguish between an anatomical structure, a functional structure, an effective structure, and an informational structure, which may be state-dependent or averaged. The anatomical structure/connectivity corresponds to the graph of the system (which element is connected to which). The functional structure/connectivity corresponds to the observed average correlations (of any order) among elements. The effective structure/connectivity captures the average causal effects of elements on other elements. Finally, the *informational connectivity* corresponds to the maximally integrated conceptual structure (retrodictive and predictive) generated by the system in a given state (or averaged over many states).

27 The complete characterization of an experience or quale would thus require specifying all of the concepts (cause-effect repertoires in Q) of a complex. From the intrinsic perspective, these concepts provide the information necessary to distinguish that experience from any other. From the extrinsic perspective, knowing these distributions and their degree of irreducibility, one would know all there is to be known about that experience. It is interesting to ask how much information that is (in terms of algorithmic complexity or incompressible information). Clearly, the input-output matrix of a system (or transition probability matrix TPM) plus the state vector, if known and available to perform manipulations (injecting noise), could be used to derive all the quantities discussed here. However, the information in the TPM is both uncompressed and implicit. A TPM is *uncompressed* if it can be reduced to the product of the smaller TPMs, as indicated by $\varphi^{MIP} = 0$. More generally, finding $_{max}\varphi^{MIP}$ and $_{max}\Phi^{MIP}$ over subsets of elements would indicate how best to compress a large TPM into the product of smaller, maximally irreducible TPMs, plus some extra terms. Also, it may turn out that a TPM at the finest spatio-temporal grain may be compressed to a coarser spatio-temporal grain with no loss (or indeed gain) in information. This aspect is captured again by finding $_{max}\Phi^{MIP}$ over different spatio-temporal scales. The TPM is also *implicit*: while it contains all the information necessary to find complexes and specify their quale, making them explicit requires work. One must extract the repertoires specified by each element and subset of elements, find the MIP to establish which subsets integrate information, which sets of elements are maximally irreducible (concepts and then complexes), and at which spatio-temporal grain size. This requires examining the effects of a large number of perturbations (by performing partitions and injecting noise/max entropy) within a large combinatorial space. At a minimum, one would need to calculate probability distributions specified by each element, from which one can calculate all the distributions specified by subsets of elements (as the product of distribution at lower levels in the power-set). From this one can establish, through appropriate partitions, which subsets specify maximally irreducible distributions (concepts) and which maximally irreducible subsets constitute complexes. It would be interesting to know if obtaining a complex and its quale (a maximally integrated conceptual structure) is equivalent to finding the most compressed model/description of the causal structure of a physical process.

28 In any case, *describing* a quale would not be the same as *being* that quale.

29 Note that in an unpredictable environment it is important not only to have a large repertoire of possible actions, but also to have many different ways of achieving the same effect, i.e. degeneracy (Tononi et al., 1999). High degeneracy implies both high effective information and high integration in the effect repertoire component of the concepts available to a complex. In general, if information integration is high, a small subset of elements within a complex should be

able to affect many other elements (pleiotropy). At the same time, many subsets of elements should be able to produce the same effect over a small subset of outputs (degeneracy).

30 In its simplified form (distribution of system states), the mismatch term is related in spirit to Bayesian perspectives and free-energy minimization, for example (Rao, 2004; Friston, 2009), as well as to the notion of causal entropy. It is also related to the quantity minimized when a Bolzmann machine learns, by changing its connections, to generate endogenously with high probability the states of the input units triggered exogenously by the environment (Ackley et al., 1985).

Note that for $_{max}\Phi^{MIP}$ to be high *on average* it is necessary that *both* causes *and* effects of any subset of neurons are reliable and selective, that is, $_{max}\varphi^{MIP}$ must also be high on average. Usually, this will go along with high (average) mutual information between past inputs and future outputs. Note also that with cause-effect information the emphasis is both on understanding (prediction) and control, since it is their minimum that is reflected in $_{max}\varphi^{MIP}$. On the input side, for example, a subset of elements would try and adapt its connections to capture correlations (coincidences) from the environment (through feed-forward connections). On the output side, it would try and adapt its connections to exert maximum control over its environment. To do so, a subset of elements must evaluate if its present actions 'make a difference'. Crucially, this can be assessed by sampling its future inputs: if different states of a subset yield specific outputs (actions) in the present, and these in turn reliably produce specific inputs (perceptions) in the future, then the subset can conclude that its actions are making a difference. This assessment can be made from the intrinsic perspective of the subset and must be causal, as it requires that the subset perturbs its environment by producing different outputs. In general, the maximization of $_{max}\Phi^{MIP}$ (and of the average cause-effect information over the channel between the subset's outputs and inputs) should take into account physical constraints associated with different subset states. In the case of neurons, for instance, the constraint that being ON is metabolically more expensive than being OFF would assign more expensive subset states (ON) to perceptions-actions that are rare (selectivity) and thus have high cause-effect information (Balduzzi and Tononi, 2012). Other physical constraints, for instance on the number of inputs and outputs, suggest that simply maximizing average cause-effect information for each element is not sufficient. One reason is that what may be optimal for individual elements is not necessarily optimal for combinations of elements and vice-versa (for example, an individual element could maximize cause-effect information by merely connecting back to itself). Moreover, it is essential that different subsets of neurons specialize in such a way that they generate different cause-effect repertoires, thus covering different portions of conceptual space (specificity) and avoiding redundancy.

A strategy neurons can use is to strengthen subsets of connections A only when a persistent feed-forward input that made them burst (primarily through AMPA receptors) is associated with a persistent feedback burst (primarily gated by NMDA receptors) on the same connections. In this way a neuron can eventually ensure that firing together with a certain subset A of synergistic neurons will produce input A, while firing together with a different subset B of synergistic neurons will produce input B, and so on. By following such a local rule, the neuron will not only optimize control (if I burst for A, I get A, if I burst for B, I get B, and so on) but promote an outcome in which different subsets of neurons generate different effects, yielding a larger number of concepts. This ensures that different subsets of elements specialize to perform different functions (maximizing information) and yet do so synergistically (maximizing integration), which is bound to be advantageous in an environment with a rich causal structure. Of course, this will also ensure both an increase in capture (thanks primarily to plasticity in feed-forward connections) and a decrease in mismatch (thanks primarily to plasticity in feed-back connections). The overall strategy should be to maximize the average value of integrated (irreducible) information proceeding in a bottom-up manner. That is, subsets of elements should maximize the cause-effect information they generate above and beyond that generated independently by their parts. Further constraints, for instance on total connection strength in the presence of noise, further suggest that subsets of elements should optimize maximally integrated (irreducible) information, so as to focus their resources on core causes and effects (at the same time, weak connections outside the core concept would provide a repertoire of alternative concepts that can be strengthened and eventually substituted as core concepts in the face of an unpredictable, changing environment).

As was mentioned above, the environment will ultimately take care of selecting those cause-effect repertoires that are not only highly effective/informative, but that match its causal structure and are therefore adaptive. A related constraint is that, in large adaptive systems such as brains, most elements connect to other elements within the system rather than directly with the environment. This has the advantage of allowing the brain's actions to be guided by memory

(intrinsic models) and thereby go far beyond the current sensory evidence (Tononi et al., 1996; Tononi and Edelman, 1997). On the other hand, in such systems it becomes important to ensure that the memory continues to match the environment by 'capturing' it. This can be done by systematically selecting in favor of the concepts generated when the system is embedded in the broader cause-effect loop that includes the external environment (wake), and against those produced by the system in isolation (sleep/dreaming). During wake, the system would adjust its connections, both feed-forward and feed-back, primarily by strengthening those that pick up correlations in the world here and now ('World' condition, as different from 'Noise'), thus increasing the capture term. Net strengthening is expected due to the bias towards assigning higher cause-effect information to ON states, which must be achieved by strengthening rather than weakening connections if ON states are to percolate throughout the brain (Balduzzi and Tononi, 2012). During sleep, the brain disconnects from the environment, and generates intrinsic activity patterns that provide a fair sampling of its overall intrinsic model of the world, not tied to the particular correlations in the environment here and now ('Dream' condition). By protecting those synapses that are most strongly activated and depressing those that are not, sleep can enforce a process of competitive down-selection. The net result is to decrease the mismatch term by favoring intrinsically generated states that match the overall statistical structure induced by the environment, and by eliminating those that do not (fantasies). An iterative process that intersperses periods of learning during wake with cycles of synaptic down-selection during sleep would increase signal to noise ratios by maintaining selectivity, favor the extraction of gist, the integration of new with old memories, and desaturate the ability to learn (Tononi and Cirelli, submitted).

31 It is worth considering simplified estimates of matching that may be applicable to neuroimaging data. For example, distance measures could be applied to covariance matrices obtained under *World*, *Noise*, and *Dream* conditions assuming a multivariate Gaussian distribution.

Also, since energy constraints in the brain force firing to be sparse and thus more informative than non-firing (Balduzzi and Tononi, 2012), one would expect that a larger number of concepts would be activated by *World* than by *Noise*. Moreover, one would expect that different inputs should lead to different responses in the *World* condition, but to the same, stereotypic response ("noise") under the *Noise* condition. Moreover, for a well-adapted brain, the mismatch term should be low, as suggested by the similarity between waking and dreaming consciousness (Nir and Tononi, 2010) and by the similarity of activity patterns in wake and sleep ('reactivation'). Thus, estimates of the difference between the complexity of spatio-temporal activation patterns under *World* and *Noise* conditions that reflect the brain's capture of *World* may also work as approximate measures of matching. Finally, since high matching M requires high $_{max}\Phi^{MIP}$, estimates of optimal matching for a given brain, being proportional to its capacity for information integration, could be used to estimate consciousness itself.

32 Since consciousness undoubtedly exists (indeed, it is the only thing whose existence is beyond doubt), if each individual consciousness is an integrated conceptual structure, then integrated information must be a fundamental ingredient of reality – as fundamental as mass, charge, or energy (Tononi, 2008).

33 Note that, in a deterministic system, there is always a one-to-one mapping of states forward in time (1 is followed by 2). However, the backward mapping can be one-to-many (2 could have been preceded by many previous states), implying that the effective information generated by the mechanism in state 2 may be insufficient to specify its previous state. In other words, going backward in time reversibly would require extra information than that available to the mechanism in its current state. Hence irreversibility.

34 A useful way of seeing this point from a neurophysiological perspective is to compare a response mediated by a conscious corticothalamic main complex with one mediated by a reflex arc. Say the task is to blink if a light turns on and not to blink if it turns off (cf. the photodiode thought experiment). For a reflex arc – say one producing a blink in response to the light – the underlying wiring diagram includes just a small chain of neurons and connections, the nodes and connections in the reflex arc (assuming, for the sake of the argument, that the nodes of the reflex are reciprocally connected). The corresponding integrated conceptual structure would be equally small – indeed just a simple concept, and it would carry hardly any experiential quality. For a conscious human performing the same task, instead, the relevant wiring diagram would be vast, including a large portion of the corticothalamic system. The corresponding integrated conceptual structure would be extraordinarily large and complex, containing a huge number of distinct concepts. This quale would correspond to the experience of seeing the light, and may also include, in a context-dependent manner, the intention to blink, or try and suppress the blink, or to say the word "light", or to interrupt the experiment, and so on. This com-

plexity may be ignored when examining how the task is performed from a limited extrinsic perspective, say that of a neurophysiologist looking for the neurons that are activated when performing the task: one may single out a causal chain 'inscribed' on top of the corticothalamic complex and represented by the neurons that fire, from a photoreceptor in the fovea to a motoneuron driving the blink, while ignoring the rest of the system. However, what is missed in such an approach is the large set of counterfactuals. In the case of the corticothalamic main complex, as opposed to the reflex arc, the silent neurons matter: if they had fired, in any of innumerable combinations, rather than having remained silent, the output would have been different. In other words, in a complex it is just as important that some neurons fire as that the others do not, whereas in the reflex arc there are no other neurons that could affect the end result. The tendency to consider that only neurons that fire 'cause' effects, or generate information, is natural enough, but it is insufficient when dealing with an integrated system. By applying perturbations to the corticothalamic system, it would become apparent that the "causal funnel" (i.e. extended receptive field or cone of influence) of a neuron of the main complex ultimately leading to a voluntary blink involves the entire main complex: in other words, its output might have been different not only if the neurons prior to it in the causal chain that had fired had instead not fired, but also if neurons that were silent had instead fired. In other words, an input – output pair (function) performed by a large integrated conceptual structure is performed *for many different reasons*. In general, it is possible to implement any given function with a simpler circuit, like the reflex arc, but in that case the function will be performed *for just one reason*. However, by exploiting counterfactuals, one can demonstrate what the integrated system has that the reflex system lacks.

35 This notion of causation can also be said to quantify IF-THEN statements (IF 1 THEN 2) while considering all possible counterfactuals (If 2-THEN…), and while examining all compound IF-THEN conditions (IF 1 AND 2, THEN…) as long as they are irreducible.

36 Seeing conscious deliberations as maximally irreducible has some relevance for the issue of free will. Consider for example the requirement for *autonomy*: to be free, one must certainly be independent from constraints outside one's deliberating consciousness. These include both environmental constraints, such as limitations that force us to a particular choice or that impede our own choice, and unconscious, 'alien' constraints that, while generated somewhere within our brain, affect our actions largely outside the control of the conscious self. Given the definition of a complex, a conscious choice is necessarily autonomous, as it is made intrinsically.

The requirement for *understanding* implies that, to be free, a choice must be based on a concept of what is at stake – for example, one can freely choose between right and wrong only if one has a notion of which actions are right and which are wrong under some circumstances. According to IIT, a complex can be held responsible for a certain choice only if it has a mechanism implementing the corresponding causal concept, in this case the backward component of the concept. For example, I must have a concept corresponding to the distinction between right and wrong (IF certain sets of past states occur, THEN certain sets of future actions/omissions are right/wrong) to be responsible for that choice – that concept is a maximally irreducible cause for my action. Similar considerations apply to the requirements for *self-control*, since the forward component of concepts within a quale ensures control.

The requirement for *irreducibility* implies that a choice can only be free if it cannot be ascribed to anything less than myself – I am the only entity that can be said to be *responsible* for my choice. That is, when asking who is responsible for the choice, the answer should be 'me', meaning *all* the circuits underlying my present conscious experience, and nothing less than that. IIT indicates that each experience is a maximally integrated conceptual structure generated by a complex, and therefore what it will choose given a particular present state cannot be ascribed to anything less than the full structure, with all its concepts (recall the light-blink example of a previous note). This structure is supremely causal to account for the choice in that it is maximally irreducible – anything less won't do, anything more won't matter. Furthermore, the choice happens at the macro-level at which Φ^{MIP} is maximized, meaning that our conscious choices are not an illusion supervening upon micro-level events that are the true causes, as is often assumed. Indeed, the macro-level exists only if it has more causal power than the micro-level, which it then supersedes. Thus, each choice is a choice of the whole complex, not reducible to a number of choices made within nearly independent modules, each in a limited context, or to choices made by micro-elements. Therefore a choice is the freer, the more it is conscious: *more consciousness, more freedom*. Moreover, a bit paradoxically, *a choice is the freer, the more it is determined* (intrinsically). This is one fundamental sense in which the key notion of *alternative possibilities* – the feeling that one could have acted otherwise, which is essential to the feeling if being responsible for one's action, is captured by a large integrated conceptual structure: such a structure implies a very large number of counterfactuals (alter-

native possibilities) that are under the control of the agent (they are part of his consciousness). In other words, a conscious choice is one in which a large number of highly informative concepts that make up my perceptions, thoughts, beliefs, desires, feelings, memories, and character, all concur in determining a choice in the integrated 'tribunal' of consciousness. Note however that, even though every conscious choice involves a large number of counterfactuals, it is still useful to distinguish between 'deep' and 'shallow' conscious decisions, based on how many concepts are dirctly involved in determining the choice here and now. At one extreme, the decision to request a divorce or not is likely to involve simulatenously many different concepts within the complex, so it is deep. At the other extreme, the decision to flex one's finger or not during an experiment on free will depends on just a small number of concepts (do I feel the urge or not), so it is shallow. This is because the previous conscious decision to participate in an experiment on free will has had the consequence of fixing most variabels within the main complex, so the only variable that is left free to vary is the 'urge' to act.

In this view, freedom requires first and foremost irreducibility, meaning that a choice cannot be ascribed to anything external, or anything less, than the agent. However, indeterminism also plays a role, though not the usual role of reducing responsibility by substituting it with chance. Recall that if a complex generates maximal integrated information at a macro-scale in space or time (say neurons instead of subatomic particles, and over hundreds of milliseconds), this means that: i) the system is most determined, in an informational/causal sense, at that macro-scale than at any micro-scale; but ii) it is also necessarily under-determined, because the macro-level can be more informative/causal than the micro-level only if there is some indeterminacy. Given that our own consciousness appears to flow at a macro spatio-temporal level, some degree of indeterminism is a given (in line with both physical sources of indeterminacy and the simple fact that the environment is unpredictable). But IIT does not consider indeterminism as a drop of randomness that instils some arbitrariness into a preordained cascade of mechanisms, thus decreasing their causative powers. Rather, in this view indeterminism provides a backdrop of ultimate unpredictability against which macro-level, integrated mechanisms fight to increase understanding and control – a fight for increasing the causative powers of consciousness, and the more these increase, the more freedom increases. But since this is a battle against a backdrop of indeterminism, its results are never completely predictable. In other words, freedom of will is a fight in which order (integrated information) tries to minimize disorder (lack of constraints) by taking into account as many constraints (knowledge) as possible. A bit like building a society or a civilization out of relative chaos, or a bit like evolution creating macro-order out of micro-level disorder, thus increasing complexity. But as with societies, civilizations, and evolution, what will actually occur can never be predicted exactly before it happens, and micro-fluctuations – a queen and a squire falling in love, two lizards separated from the mainland after a flood – may initiate an extraordinary turn of events that nobody could predict, not even the universe itself.

37 It is interesting to consider how the notion of maximally irreducible set of past causes of future effects maps onto accounts of trajectories of dynamical systems, for example accounts of how an element may be enslaved by one of two weakly coupled attractors, though being subjected to causal influences from both. It is also interesting to consider how the intrinsic notion of causation indicated here maps onto an extrinsic notion of causation developed along parallel lines (Hoel et al., in preparation). In the intrinsic perspective, one takes a mechanism in a state and asks what could have potentially caused it and what are its potential effects (*causal power*). From the extrinsic perspective, one takes a given event (i.e. an observed state) and considers what past event actually caused it and what are its actual future effects. In this way, it is possible to define an extrinsic notion of *causal action* based on the sufficiency (reliability) and necessity (selectivity) of the transition from one event to the next, and the size of the repertoire of counterfactuals. By applying the information, integration, and exclusion postulates, one can then proceed to partitions to identify maximally irreducible ("core") cause-effects as well as sets of cause-effects ("cause-effect complexes").

38 Note that exclusion can also be applied to define not only the set of elements and spatio-temporal grain at which information integration reaches a maximum, but even the mechanism itself, considered as the maximally irreducible input-output function among a set of possible functions. For example, consider a potential mechanism "neuron" that is firing. Why should every extra-synaptic glutamate receptor on the membrane of the neuron not be considered as part of the neuron's mechanism, since it is capable of binding stray glutamate molecules and produce some slight depolarization? Because by exclusion, after "freezing" the boundary conditions, the contribution of that receptor to the neuron's firing, unlike the contribution of the strongest synapses on the neuron, can be eliminated (reduced) with minimal informational/causal loss. In this way, identifying the core cause of the neuron's

firing reveals the neuron's core mechanism – its most irreducible causal role.

39 This is because concepts are ways of grouping possible states of the elements of a complex (for a catalogue of Boolean concepts over systems of up to 4 binary elements, see Feldman, 2003). The number of concepts is even larger if one considers cause-effect pairs, as in the present case, not to mention sequences of states over time (*melodies*). It is worth noting that the number of concepts over a system of elements that can be implemented by the elements themselves (through causal interactions) is therefore only a fraction of all possible concepts that could be formulated over the system. That is, a system's understanding/control of itself is necessarily incomplete, an instance of incompleteness that is reminiscent of well-known incompleteness theorems in axiomatic systems. On the other hand, while it is conceivable to implement a much larger number of concepts to categorize and control a system externally, by resorting to additional, external elements (from the extrinsic perspective), these concepts would not be integrated within a single experience (unless they are in the mind of the beholder). So, while even a system that maximizes the number of concepts within a single maximally integrated conceptual structure would remain substantially incomplete, it would nevertheless have the benefit of much more comprehensive understanding and control, since its concepts would be experienced 'together'.

40 With due differences, a similar strategy is employed in relational databases and in object-oriented programming.

41 Note that neural oscillations and fine temporal synchronization among units responding to different aspects of a particular object or event (specifying related concepts) can greatly enhance these differential effects. This is because neurons that are oscillating in phase or are otherwise synchronous can have greater joint (synergistic and thus irreducible) effects on downstream targets, due to the short integration time constants of membranes, various receptors and so on. See for example (Sporns et al., 1991; Tononi, 1992; Tononi et al., 1992a; 1992b; Buzsáki 2010).

42 In this context, language has obvious benefits. Not only does language add many new concepts, including concepts specified over internal system units (as opposed to input units), but it also helps making many concepts more accessible.

43 Presumably, the underlying neural mechanisms would be similar to those that are thought to mediate attentional effects, including back-connections and lateral connections, the action of the reticular thalamic nucleus, the precise timing of inhibition, and the multiplicative properties of NMDA receptors (Tononi et al., 1992a; Roelfsema and van Ooyen, 2005). The latter seem especially well suited to ensuring that a neuron's level/duration of firing/synchronization can be modulated by 'contextual' input conveyed for example by back-connections, while at the same time preserving the selectivity of the main concept specified by that neuron, as typically conveyed by its driving input through forward connections.

44 As discussed elsewhere (Tononi, 2010), certain concepts (questions and answers) are more easily accessible than others. One can speculate that concepts would be more accessible or 'explicit' if the relevant neural units can themselves be accessed easily and selectively, for example through back-connections, which may vary depending on the cortical area, whether they are locally clustered rather than widely distributed, and so on. Ease of access may be partly related to the distinction between access and phenomenal consciousness (Block, 2005; Tononi; 2010). Alternatively, one could say that information that is specified by a concept (a maximally irreducible probability distribution specified by a subset of elements) is explicit, whereas information that can be inferred but is not specified by a concept remains implicit. For example, with respect to the cause repertoire, if C copies A and D is an XOR of A,B, and both C and D are OFF, then one can infer that B is OFF, but there is no explicit concept for it.

45 However, the possibility of a non-dominant complex of relatively high Φ that may be largely excluded from access to behavioral outputs (a "minor" complex) should not be discounted, for example in conditions of perceptual rivalry.

46 While a precise evaluation of Φ and Q for realistic systems is computationally prohibitive, by making certain plausible assumptions it should be possible to develop practical measures that reflect the degree of information integration in different systems or in the same system under different conditions. In addition to the approaches suggested in Note 31, one could evaluate cause-effect information (or, making further assumptions, mutual information) for individual elements with respect to the entire system at previous and following time intervals, then for subsets of 2 elements, 3 elements and so on, up to the full system with respect to itself. In the extreme case of a completely homogeneous systems (full connectivity), the expected value of cause-effect information would not increase with subset size (knowing the state of 2, 3 or more elements would add no information compared to knowing the state of 1 element); in a completely modular systems (say, a system made of disconnected modules of 2 elements), the expected value of cause-effect information should grow linearly with subset size; in a complex system characterized by both

functional segregation and integration, cause-effect information can grow supralinearly, where the supralinear increase should reflect the points in qualia space having $\varphi > 0$ (knowing the state of more elements would add information compared to the information provided by individual elements taken separately). Ideally, one would evaluate these quantities under maximum entropy perturbations and, for each subset of elements, consider the increase in cause-effect information over its minimum information partition. In practice, one may resort to observed probability distributions and, for representative subsets of elements, evaluate the increase (synergy) or decrease (redundancy) of information compared to the summed information of the subset's constituting elements. Also, approximate measures of matching could be used to estimate integrated information, and by extension consciousness, both across individuals and across species, since the maximum value of matching for a given brain is likely to be limited by its value of Φ. Such an approach may be particularly useful when dealing with pathological conditions, both during development and after brain lesions.

References

Ackley D.H., Hinton G.E., Sejnowski T.J. A learning algorithm for Boltzmann machines. *Cognitive Science*, **9**: 147-149, 1985.

Ay N. and Polani D. Information flows in causal networks. *Systems Research*, 1-15, 2006.

Balduzzi D. and Tononi G. Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput. Biol.*, **4**, p.e1000091, 2008.

Balduzzi D. and Tononi G. Qualia: the geometry of integrated information. *PLoS Comput. Biol.*, **5**, e1000462, 2009.

Barrett A.B. and Seth A.K. Practical measures of integrated information for time-series data. *PLoS Comput. Biol.*, **7**, e1001052, 2011.

Bateson G. *Steps to an ecology of mind*. Chicago, IL, University of Chicago Press, 1972.

Beggs J.M. The criticality hypothesis: how local cortical networks might optimize information processing. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, **366**: 329-343, 2008.

Bennett C.H., Gacs P., Li M., Vitany P.M.B., Zurek W.H., Information distance. IEEE *Trans. Information Theory*, **44**: 1407-1423, 1998.

Braun J. and Mattia M. Attractors and noise: Twin drivers of decisions and multistability. *NeuroImage*, **52**: 740-751, 2010.

Buzsáki G. Neural syntax: cell assemblies, synapsembles, and readers. *Neuron.*, **68**: 362-385, 2010.

Bělohlavek R. Representation of Concept Lattices by Bidirectional Associative Memories. *Neural. Comput.*, **12**: 2279-2290, 2000.

Chen W. et al. A few strong connections: optimizing information retention in neuronal avalanches. *BMC Neurosci.*, **11**: 3, 2010.

Creutzig F., Globerson A., Tishby N. Past-future information bottleneck in dynamical systems. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **79**: 1-5, 2009.

Crick F. and Koch C. A framework for consciousness. *Nat. Neurosci.*, **6**: 119-126, 2003.

Deco G., Rolls E.T., Romo R. Stochastic dynamics as a principle of brain function. *Progr. Neurobiol.*, **88**: 1-16, 2009.

Deco G., Jirsa V.K., McIntosh A.R. Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nat. Rev. Neurosci.*, **12**: 43-56, 2011.

Edelman G.M. *Neural Darwinism – the theory of neuronal group selection*. New York, Basic Books, 1987.

Feldman J. A catalog of Boolean concepts. *J. Math. Psychol.*, **47**: 75-89, 2003.

Friston K.J. Transients, metastability, and neuronal dynamics. *NeuroImage*, **5**: 164-171, 1997.

Friston K.J. The labile brain. I. Neuronal transients and nonlinear coupling. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **355**: 215-236, 2000.

Friston K. The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.*, **13**: 293-301, 2009.

Gazzaniga M.S. Forty-five years of split-brain research and still going strong. *Nat. Re. Neurosci.*, **6**: 653-659, 2005.

Gros C. Cognitive computation with autonomously active neural networks: an emerging field. *Cogn. Comput.*, **1**: 77-99, 2009.

Hutter M. *Universal artificial intelligence: sequential decisions based on algorithmic …* Springer, 2005.

Joshi S., et al., 2011. Comparing Distributions and Shapes using the Kernel Distance. *Proceedings of the 27th annual ACM symposium on Computational geometry*: 47-56, 2011.

Koch C. *The quest for consciousness: a neurobiological approach*, Denver, CO, Roberts and Co, 2004.

Koch C. and Tononi G. Can machines be conscious? *Ieee Spectrum*, **45**: 54-59, 2008.

Lehmann D., Pascual-Marqui R., Michel C. EEG microstates. *Scholarpedia*, **4**: 7632, 2009.

Musso F., et al. Spontaneous brain activity and EEG microstates. A novel EEG/fMRI analysis approach to explore resting-state networks. *NeuroImage*, **52**: 1149-1161, 2010.

Nir Y. and Tononi G. Dreaming and the brain: from phenomenology to neurophysiology. *Trends Cogn. Sci.*, **14**: 88-100, 2010.

Rabinovich M., et al. Dynamical principles in neuroscience. *Re. Modern Phys.* **78**: 1213-1265, 2006.

Rao R.P.N. Bayesian Computation in Recurrent Neural Circuits. *Neur. Comput.*, **16**: 1-38, 2004.

Roelfsema P.R. and van Ooyen A. Attention-gated reinforcement learning of internal representations for classification. *Neur. Comput.*, **17**: 2176-2214 2005.

Singer W. Distributed processing and temporal codes in neuronal networks. *Cogn. Neurodyn.*, **3**: 189-196, 2009.

Solomonoff R.J. A formal theory of inductive inference. *Information and Control*, **7**: 224-254, 1964.

Sporns O., Tononi G., Edelman G.M. Modeling perceptual grouping and figure-ground segregation by means of active reentrant connections. *Proc. Natl. Acad. Sci. U S A*, **88**: 129-133, 1991.

Sporns O. *Networks of the Brain*, MIT Press, 2010.

Tononi G. Reentry and the Integration of Brain Function. *Comunicazioni Scientifiche di Psicologia Generale*, 1992.

Tononi G., Sporns O., Edelman, G.M. Reentry and the problem of integrating multiple cortical areas: simulation of dynamic integration in the visual system. *Cereb. Cortex*, **2**: 310-335, 1992a.

Tononi G., Sporns O., Edelman G.M. The problem of neural integration: induced rhythms and short-term correlations. In: Basar E. and Bullock T. (Eds.). *Induced rhythms in the brain*. Boston, MA, Birkhäuser: 367-395, 1992b.

Tononi G., Sporns O., Edelman G.M. A complexity measure for selective matching of signals by the brain. *Proc. Natl. Acad. Sci. U S A*, **93**: 3422-3427, 1996.

Tononi G. and Edelman G.M. Information: In the stimulus or in the context? *Behav. Brain Sci.*, **20**: 698, 1997.

Tononi G., Sporns O., Edelman G.M. Measures of degeneracy and redundancy in biological networks. *Proc. Natl. Acad. Sci. U S A*, **96**: 3257-3262, 1999.

Tononi G. An information integration theory of consciousness. *BMC Neurosci.*, **5**: 42, 2004.

Tononi G. Consciousness as integrated information: a provisional manifesto. *Biol. Bul.*, **215**: 216-242 2008.

Tononi G. and Laureys S. The Neurology of Consciousness: An Overview. In: Laureys S. and Tononi G. (Eds.). *The Neurology of Consciousness: Cognitive Neuroscience and Neuropathology*. Academic Press, Elsevier: 375-412, 2009.

Tononi G. Information integration: its relevance to brain function and consciousness. *Arch. Ital. Biol.*, **148**: 299-322, 2010.

Van de Ville D., Britz J., Michel C.M. EEG microstate sequences in healthy humans at rest reveal scale-free dynamics. *Proc. Natl. Acad. Sci. U S A*, **107**: 18179-18184, 2010.

Wennekers T. On the Natural Hierarchical Composition of Cliques in Cell Assemblies. *Cogn. Comput.*, **1**: 128-138, 2009.